May-June 2020
ISSN: 0193-4120 Page No. 9328 - 9333

Iterative Text-to-Image conversion using Recurrent Generative Adversarial Tell, Draw, Undo, Repeat

Vinay Varma Nadimpalli, Venu Vardhan Reddy Tekula,
Sumanth Javvaji, Sedimbi Satya Pramod, Jyothisha J. NairDepartment of Computer Science and Engineering, Amrita Vishwa Vidyapeetham,Amritapuri Campus, India.

Article Info
Volume 83
Page Number: 9328 - 9333
Publication Issue:
May - June 2020

Article History
Article Received*: 19 November 2019*
Revised*: 27 January 2020*
Accepted*: 24 February 2020*
Publication*: 18 May 2020*

Abstract:
Conditional text-to-image is an active area of research. We propose a novel Recurrent GAN model architecture that can generate 2-D images from input text in an iterative manner. This is different from the one-step text-to-image generation as the model will be given continuous instructions carrying information on how to modify the most recently generated image. To generate an image in the current time step, the model would take all the previous instructions up to the current time step and will generate the image from the previous time step. We propose a novel method, which is the ability of the model to undo the modifications that it has done to the image on the previous time step. This is very essential in situations when the text instruction given to the model is incorrect (possibly unintentional human errors). The aim of the proposed model is to march towards a complete capacity of interactive text-to-image generation and to enhance the user experience from a Human-Computer Interaction (HCI) perspective.

Keywords*: Computer Vision, Human-Computer Interaction, Iterative Text-to-Image generation, Natural Language Processing, Recurrent Generative Adversarial Networks.*

INTRODUCTION

Communication has played a vital role in connecting people right from the early days of human civilization. It has greatly contributed and continues to contribute to the progress of society. Humans have always chased the ability to perfectly articulate their opinions, stories, emotions, etc. Research in neuroscience for the art of presentation tells us that visual depiction of data/information is more pleasing to the brain and makes it easier to process and remember.

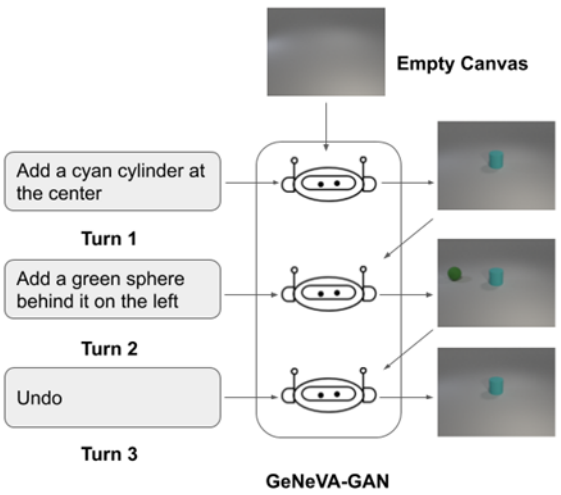

Figure 1: An illustration of how the undo feature works.

May-June 2020
ISSN: 0193-4120 Page No. 9328 - 9333

A lot of research has been put into the one-step text-to-image generation in recent years [1]–[7]. Although all of these models produced exceptionally high-quality images, they lack the ability to make modifications to the generated images.. Such a system is necessary in modern times where diseases like Aphasia and Dyslexia are on the rise and stand as a barrier for communication.

Such a model was proposed by Alaaeldin et al. [9] which can take in continuous linguistic input and can iteratively generate images. The main objective of the model we propose is the ability to revoke the changes which were applied in the previous time step, in case the input text has incorrect instructions. We present an example for this task in Figure 1. This is vital to achieve a truly interactive image generation experience.

PROBLEM DEFINITION

Vision is the medium through which humans mostly interact. Machines that can understand the world and generate new images and videos have wider applications in pioneering interactive education, multimedia creation, and creative arts. Visual representation of data makes comprehension easier for the mentally disabled, the deaf and people suffering with various other disabilities that hinder the brain's interpretation capacity. Also, language is an obstacle when seeking to communicate with people across the world. Though there are many neural machine translation models that solve this issue, they are limited to only a few languages across the globe. The data required to train a translation model for regional languages is quite insufficient. Hence having a model that can convert human language with complex linguistic rules into images can solve the issue since images are a universal way of representing data.

Continuous generation of images will revolutionize the multimedia industry and will pave the way for the rapid creation of content for professional artists. It will help them exactly elucidate their thoughts. There has been some work on converting text to image, however, these models are not interactive as these models cannot fix any mistakes on the generated image which are visible to the user. This feature of improving the interactiveness will further help the user to steer the model to draw what the user actually needs.

This type of model requires a dataset where each sample in the dataset is a consecutive sequence of image-label pairs with each image-label pair directly dependent on the previous image-label pairs. iCLEVR dataset, the iterative version of CLEVR dataset has been used to train the model. The iCLEVR dataset consists of 10,000 sequences, each sequence having 5 images with their corresponding text. That makes it a total of 50,000 image and instruction pairs.

The adversarial model consists of a Generator and a Discriminator. The Generator's task is to take in a noise image (a Gaussian distribution) and the text from the current time step encoded by a Gated Recurrent Network (GRU) and previous instructions encoded by a Bi-Directional Gated Recurrent Network. The Discriminator's job is to take in the ground truth image for the text and then calculate a measure r-sim, the similarity between the generated image and ground truth image. The Discriminator then gives a loss value based on the r-sim value. This loss value is later fed into the Generator which adjusts its weights so that the loss value at the Discriminator increases. The goal is to minimize the loss of the Generator so that it can pass the Discriminator test. Thus paving the way to building a Recurrent GAN model that can draw upon the previously generated images.

RELATED WORK

A lot of research across the world has been conducted to model a machine that takes user input as text, interprets it, and then converts that information from the text embedding space to a 2-D image space. Variational Autoencoders (VAEs) [17] gained popularity by maximizing the lower bound of data likelihood. This was followed by the rise of autoregressive models (like PixelRNN [19]) that modelled the conditional distribution of pixel space.

May-June 2020
ISSN: 0193-4120 Page No. 9328 - 9333

Generative Adversarial Networks (GANs) [14] have greatly propelled conditional image generation. Visual Question Answering VQA [18] solved the task of modelling a machine which can answer natural language questions about a given image by using Reed et al. [15] pioneered using GANs for text-to-image synthesis. With the introduction of StackGAN [3], the process was extended to a two stage procedure. It had two pairs of generators and discriminators, where the first GAN produced a low resolution image and was passed to the next GAN in connection to produce a high resolution image. This idea was further enhanced by StackGAN++ [16] by introducing the concept of conditioning augmentation within the concept of StackGANS, resulting in a higher resolution image. Sharma et al. [12] researched a model named ChatPainter, which is non-iterative, but was conditioned on MS COCO and an RNN encoded dialogue which were similar to captions instead of directly using the captions. Mansimov et al. [20] explored the use of attention mechanism for text inputs during the process of text-to-image synthesis.

Later, AttnGAN [7] was introduced which explored the idea of using attention locally additional to attention for the entire sentence. Yang et al. [13] proposed the recursive generation of images, which was unsupervised in nature, where the background is generated first and then it is subsequently conditioned to further generate the foreground and the mask, which are later combined using an affine transformation. Generative Adversarial models such as StackGAN [3] and AttnGAN [7] have been able to generate high-resolution images based on the input text given by the user.

The drawbacks of these models which we plan to address are

- Accept feedback from the user and generate new images accordingly.
- Undo a particular modification.

PROPOSED SYSTEM

The proposed system is a Recurrent General Adversarial Network with an ability to undo the modifications in any time step. We use the GeNeVA (Generative Neural Visual Artist) task discussed by Alaaeldin et al. [9] and tweak it to incorporate the idea of undoing a modification. The Figure 2 shows the architecture of the model. This task has two players, Teller and Drawer. The Teller tells the instructions to the Drawer and the Drawer tries to draw what the Teller has described. First, an empty canvas will be given to the Drawer and then, iteratively Teller and Drawer will make their actions leading the way to a continuous Text-to-Image Generation. When the Teller gives ‘undo’ as an instruction, the Drawer removes the context of the previous instruction from the Teller history and revokes the modifications from the image in the previous time step. Here the Teller is modelled as the input text and the Drawer is modelled as the recurrent GAN.

There are other components such as an Image Encoder and a Sentence Encoder which helps the flow of data from the Teller to the Drawer and back to the Teller. The Image Encoder consists of two blocks: Residual Up block and Residual Down block. The Residual Down block takes an input image and extracts the features from it, thus down-sampling the image. The Residual Up block takes in a lower-dimensional image and up-samples it into the size of the original image.

Our Generator is an Image Encoder with Residual Up Block (a ResNet with residual convolution network blocks). It takes in three inputs. A noise image, features extracted from the base image, complete context aware feature vector capturing the semantic meaning of all the instructions given till the current time step. The input text at the current time step will be sent into a Bidirectional Gated Recurrent Network (Bi-GRU, also modelled as the Sentence Encoder). The words from the input text are converted to vectors with the help of GloVe [11] embeddings which is on the top layer of the Sentence Encoder. The output from this

May-June 2020
ISSN: 0193-4120 Page No. 9328 - 9333

Sentence Encoder is again fed into another GRU to get the overall context-aware vector. This context-aware vector, before being sent into the Generator along with noise image and image feature vector, has to undergo conditioning augmentation (modelled as a simple neural network with conditional batch normalisation). With the help of this conditional augmented sentence feature, noise image, and the extracted image features, the Generator will be able to generate a new image which contains the modifications instructed by the input text.

The Discriminator takes in the generated image and the ground truth image from the previous time step. It also takes the context-aware vector which contains the semantic meaning of the history of previous instructions. This helps the Discriminator to see how relevant the generated image is with respect to the given instruction and previous image.

It will give a measure of how good the modifications are. In addition to the quality and resolution of the generated image, it is important to see if all the objects described in the input text are placed according to the user instruction or not.

For this purpose, an auxiliary objective is added to the Discriminator, which is an object detector and also an object location identifier. This helps the Discriminator compare the objects in the generated image from the previous time step and objects in the generated image from the current time step, thus telling if the number of objects and types of objects has been placed correctly or not. The object location identifier estimates the position of an object in an image. It tells if the Generator was able to capture the relevant position information of objects from the input text and place it accordingly in the image efficiently or not. This object detector and location identifier are based on the Inception-V3 model [10].

Based on the results from the object localizer/object detector, a scene graph can be estimated for both the generated image and the ground truth image at each time step. In this scene graph, the nodes are the objects identified and the edges represent the positional relationships between the objects. It is easier to compare these graphs and draw out the relationship similarity between the two images, thus helping in the assessment of the ability of the generator to modify the images in accordance with the captured context from the input text.

The Undo is one feature which can be implemented in the project. The aim of the undo feature is to allow the user to revert any change which he added in the previous iteration. Whenever you give any wrong instruction, you need to give “Undo” as your next instruction to revert back to the previous image.

At each step, an image is created from the old image and the new instruction. The features are extracted from the old image and the semantic meaning of the new instruction is added to it which is then later converted into an image which is the new image. Whenever you give “Undo” as the instruction, the semantic meaning of the previous instruction is subtracted from the new image, which then is converted to the old image.

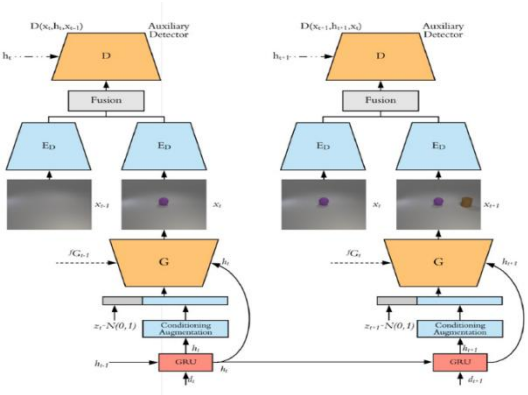

Figure 2: Model Architecture

To summarise the process step by step:

1.The base image or empty canvas (in case of the first step)/previous generated image is sent to the image encoder Res Down block to extract the features from it.

2.The input text is passed through a sentence encoder which is a Bi-directional GRU that uses the GLoVe embeddings [11].

3.The output from the sentence encoder is sent to another GRU along with the context-aware vector of all the previous instructions. This will

May-June 2020
ISSN: 0193-4120 Page No. 9328 - 9333

result in an overall context vector up to the current state.

4.This GRU produces the overall context aware sentence feature vector containing the semantic meaning of all the instructions given till the current time step.

5.The Generator takes in the concatenation of the noise vector, context-aware vector and the extracted features from the first step and generates an image.

6.Both the generated image of the current time step and the ground truth image from the previous time step is sent to the image encoder (Res down) block to extract the features. Both these features are fused together in terms of element-wise subtraction.

7.The Discriminator takes in the fused vector (fusion happens by means of element wise subtraction of feature vectors extracted from generated image and ground truth image) and also the context-aware network. The auxiliary object detector and localizer helps in generating a scene graph from both the generated image from the current time step and ground truth image from the previous time step.

8.R-sim is a metric used to assess the Generator's performance. It is defined as

Here, Egt is the set of relations of ground truth image and Egen is the set of relations of the generated image.

9.This repeats for 5-time steps, (iCLEVR dataset has 5 image-text pairs in each sample) and finally an image is generated according to the user instruction.

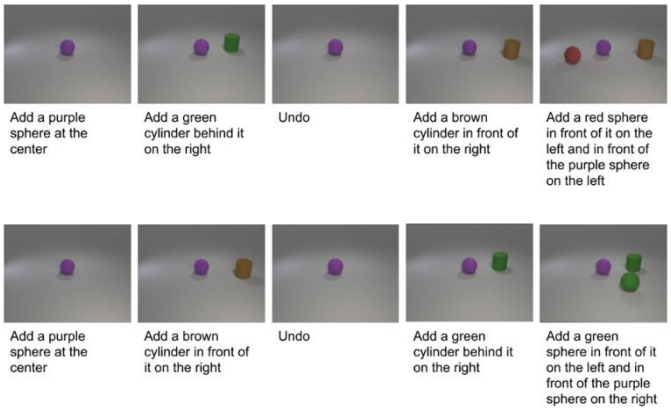

Figure 3 and 4: Results

RESULTS

The public dataset, iCLEVR is used to train the model. The dataset contains raw image and text pairs along with their positional information. The raw images undergo pre-processing (cropping and resizing) and are stored in HDF5 files (Hierarchical Data Format) so that we can feed the data as a whole to the image encoder and thereafter to our Recurrent GAN model.

The Undo feature will allow the user to revert any change which was added in the previous instruction. Figure 3 & 4 shows the results. When an instruction is given, the semantic meaning of the instruction is added to the image and a new image is generated. Whenever the user encounters an instruction which says "Undo", the semantic meaning of the recent instruction will be removed from the image and the previous image will be restored.

6 Conclusion

The proposed recurrent GAN model can generate reasonable images in the context of the provided instructions. Relational Similarity is used as a metric to assess the performance of the Generator in modifying the image from the previous time step with the context captured from the text input. We

May-June 2020
ISSN: 0193-4120 Page No. 9328 - 9333

verified that the model is able to revoke the modifications that it has done when the user has given the undo instruction to the model. Thereby, enhancing the user experience during the course of the interactive generation. The model should strictly outperform the normal non-iterative baseline model. Once the model is trained on various datasets, it would give the user the power to express their thoughts invariably.

References

1. S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text-to-image synthesis" in ICML, 2016.
2. S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw" in NIPS, 2016.
3. H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks" in ICCV, 2017.
4. S. Reed, Z. Akata, B. Schiele, and H. Lee, "Learning deep representations of fine-grained visual descriptions" in CVPR, 2016.
5. K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention" in ICML, 2015.
6. Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering" in CVPR, 2016.
7. A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, "VQA: visual question answering" in IJCV, 2017.
8. Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He, "AttnGAN: Fine-grained Text to Image Generation with Attentional Generative Adversarial Networks".
9. A. El-Nouby, S. Sharma, H. Schulz, D. Hjelm, L. El Asri, S. E. Kahou, Y. Bengio, and G. W. Taylor, "Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction" in International Conference on Computer Vision, 2019.
10. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna, "Rethinking the Inception Architecture for Computer Vision" in Computer Vision and Pattern Recognition (CVPR) 2016.
11. Jeffrey Pennington, Richard Socher, and Christopher D. Manning, "GloVe: Global Vectors for Word Representation" in Empirical Methods in Natural Language Processing (EMNLP) 2014.
12. Shikhar Sharma, Dendi Suhubdy, Vincent Michalski, Samira Ebrahimi Kahou, and Yoshua Bengio, "Chatpainter: Improving text to image generation using dialogue," in the International Conference on Learning Representations (ICLR) Workshop, 2018.
13. Jianwei Yang, Anitha Kannan, Dhruv Batra, and Devi Parikh, "LR-GAN: Layered recursive generative adversarial networks for image generation," in International Conference on Learning Representations (ICLR), 2017.
14. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in Advances in Neural Information Processing Systems (NIPS) 27, 2014.
15. Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee, "Generative adversarial text to image synthesis," in International Conference on Machine Learning (ICML), 2016.
16. 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" in Computer Vision and Pattern Recognition (CVPR), 2018.
17. D. P. Kingma and M. Welling. "Auto-encoding variational bayes". In ICLR, 2014.
18. A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra. "VQA: visual question answering" IJCV.
19. A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. "Pixel recurrent neural networks" in ICML, 2016.
20. Elman MansimovEmilio, Parisotto, Jimmy Lei Ba, Ruslan Salakhutdinov. "Generating Images from Captions with Attention" in ICLR 2016.about