RESEARCH ARTICLE OPEN ACCESS

Text to Image Synthesis in Generative Adversarial Networks

Afreen Bhumgara*, Anand Pitale**
**(Department of Computer Science, Nowrosjee Wadia College, Savitribai Phule Pune University, India*
*** (Department of Computer Science, Nowrosjee Wadia College, Savitribai Phule Pune University, India*
Corresponding Author: Afreen Bhumgara

ABSTRACT : One of the core applications of conditional generative models is to generate images from text (natural languages). In addition to running tests on our capabilities of conditional modeling, dimensional distribution at the high level; synthesis involving text to an image has numerous practical applications. Some of the applications include the creation of machine-aided content and editing photos. In the past, huge strides of progress in generative adversarial neural nets have been made. Text to image synthesis is among one of the most interesting discoveries made in the artificial intelligence field of our century. In the year 2016, the generative adversarial network for text to image synthesis was not able to generate accurate and clear images. With the advancement made in technology and adjustments to the model, it's now possible to generate clear and almost fully accurate images based on the description provided. Visualizing a scene given a detailed description is an undertaking that human beings do with less effort involved -nonetheless, it is a complex task that requires a combination of various concepts specified in natural so as to compare to how they look in real life. In this paper, we study previous work on image synthesis from text descriptions following the advances in generative adversarial networks (GANs), and experiment with better training techniques like feature matching, smooth labeling, and mini-batch discrimination.

Keywords-convolutional neural networks, GAN, generative adversarial networks, learning (artificial intelligence), machine learning, natural language processing, neural nets, object detection, text-to-image translation,

---

Date Of Submission: 16-01-2019 Date Of Acceptance:28-01-2019

I. INTRODUCTION

This section introduces us to the research study. It consists of background of the research, significance of the research, research rationale, scope, questions, hypotheses, aim, and objectives.

A. BACKGROUND OF THE RESEARCH -

Generation of an image from a text description has been a challenging issue in computer vision and natural language processing. Extraction of descriptions and attributes from an image have been also a challenging topic, it's the reverse of the later. This kind of problems is similar to the language translation problems from a high-level perspective. Text and images are two different 'languages', this can be compared to the same way similar semantics can be reduced to two languages. The difference with this kind of problem (image to text synthesis and vice versa) is that it's greatly multimodal (involves more than one model to complete the synthesis and produce output). In a semantic case, when translating a sentence such as 'this is a handsome black man' in another language such as German, there will be only minimal valid sentences equating to the original one (Zhang, Huang, Wang, & Metaxas, 2016). However, when a human being or even by use of a computer, an individual tries to produce an image matching the given description, there will be a wide range of possible solutions. These kind of problems where many outputs correspond to multiple possible solutions is known as multimodal behavior, though this problem may be easier when the error associated with it is minimized to the point that the language at most times is sequential, in that their structure is conditioned such that the incoming generation of words depends on the words previously produced. This makes synthesis of an image to text synthesis simpler than text to image synthesis due to the above condition. There is a wide scope of applications of the text to image synthesis especially when it will be apt for commercial use. For instance, one would spend less time describing an equipment such as furniture a person wants custom made rather than spending hours and sometimes days finding a similar furniture design.

B. SIGNIFICANCE OF THE RESEARCH

It is proved that Generative adversarial network (GAN) is very effective model/method. It is very effective for training Generative models. Key challenges in multimodal learning

include learning a shared representation across modalities, and to predict missing data. For the generator network module generative adversarial networks have also benefited from convolutional decoder network. (Denton, 2015) used a Laplacian pyramid of adversarial generator and discriminators to synthesize images at multiple resolutions. This work generated compelling high-resolution images and could also condition on class labels for controllable generation. (Radford, 2016) used a standard convolutional decoder, but developed a highly effective and stable architecture incorporating batch normalization to achieve striking image synthesis results. The main distinction of our work from the conditional GANs described above is that our model conditions on text descriptions instead of class labels.

The generic adversarial networks can be used for many purposes but in our research we used it for text to image synthesis where we used an text for generating an image. For better quality of images we used stack GANs which on the first stage produced an image with low resolution. The common method of training the generative model is to feed in the image along with the text and further it is kept as an pair. So by this if we give a text as an input it will produce an corresponding image for display.
In contemporary work (Mansimov, 2016) generated images from text captions, using a variational recurrent autoencoder with attention to paint the image in multiple steps, similar to DRAW (Gregor, 2015). Impressively, the model can perform reasonable synthesis of completely, novel(unlikely for a human to write) text such as "a stop sign is flying in blue skies". suggesting that it does not simply memorize. While the results are encouraging, the problem is highly challenging and the generated images are not yet realistic, i.e., mistakable for real. Our model can in many cases generate visually-plausible images conditioned on text, and is also distinct in that our entire model is a GAN, rather only using GAN for post-processing.

Generative adversarial networks (GANs) basically consist of two objects. The generator g and the discriminator d and they both act like players in a game. The idea behind is that the generator will create an image and the discriminator tries to find if the image generated is the real image or not. But the generator here tries to fool the discriminator.

II. RESEARCH METHODOLOGY

This section gives an idea about the methodology and the techniques for this research and the objectives considered most appropriate for this study. It is made up of the research purpose, study design, ethical considerations, and challenges encountered in the course of the research.

A. RESEARCH PURPOSE

The research is a theoretical assessment with an aim of providing brief descriptions about the outcomes of various GAN architectures as well as an experimental setup. For this, a detailed empirical and literature review of various models and techniques is performed that generalizes and covers various topics ranging in the field of adversarial networks. A descriptive research based on sufficient knowledge of the researcher that collaborates with the hypotheses is conducted, and the research questions set forth are investigated. Secondary data based on sources such as peer-reviewed articles, journals and study-cases is considered. Derivations and analysis are used as the basis of the research.

B. STUDY DESIGN

The research will make use of qualitative and research methods to examine the limitations and constraints by analyzing most recent peered reviewed scholarly journals and articles. The researcher performed literature review involving a survey and literature review in academic from existing online and offline libraries to select the latest journals in the topic under consideration. Furthermore, a case study analysis approach was employed in investigating the factors. Sampling for the case studies considered a selection of samples again, sampling being done by objective sampling.
Data analysis consisted of functions affecting qualitative & quantitative analysis covering the factors influencing adversarial networks through processes in order to draw an experimental design, quantitative equations and analysis and variables for measurements. Analysis of the literature review provides the past, current, and emerging trends in GANs, and therefore is significant for provoking further research. An offline self-conducted research including study methods based upon different variables as follows based on the empirical review and research. The research also uses a descriptive and correlation design-based research methodology. Modeling is used so as to substantiate the assertions.

C. ETHICAL CONSIDERATIONS

The research followed the ethic requirements. Grants and permissions to proceed with the study was obtained from all the relevant sources and authorities. The studies considered for literature and empirical review met the ethical considerations and were published in the domain of public use and interest. All the authors of the literature review considered on the basis of secondary sources and for analysis purposes are accredited in both the in-text citation and under the references section so as to adhere to the policies against plagiarism.

D. CHALLENGES ENCOUNTERED

Multiple challenges were encountered in the course of the research. Considering that it is a research based upon both the outcome of the empirical findings as well as an experimental setup, the theoretical modeling contrasted at various points and on numerous variables against the different literature reviews. As there is limited knowledge of generative adversarial networks, it was difficult to collect sufficient peer reviewed papers necessary for analysis.

III. LITERATURE REVIEW

The author has provided only two stages of synthesizing images from their text description. In the research paper, the author discovered that the text description should contain most of the image visual details. The synthesizer used these details in the generation of the image and the more the description of the image was precise and accurate the more the image output was, hence more realistic. End to end differentiable architecture that was conditioned on a comprehensive text descriptions rather than the conditioning it on a single class label. The text features were encoded using the hybrid character level convolutional neural networks which were then used to condition the deep convolutional generative model (Radford, Metz, Chintala, 2015). The image produced plausible 64by64 images though they never appeared real from a human perspective. Rather than generating the images in a single step the next author made use of two stages. Each stage has its role to play in the generation process. The author describes the two sub-domains of generation stages as first stage and second stage. In the first stage, the author used a generative model to produce rough shapes of the objects and used the sampled noise vector sampled from the prior distribution to develop a rough ideal image of the background. In the first stage, the stage description is the only input (St-Yves & Naselaris, 2017). However when we move to the second stage both the results in the first stage and the text description are used as the inputs. The inputs are used to make the roughly sketched shapes more detailed as possible and adding on the resolution of the images. The images produced in the second stage according to the author are more real and of high resolution. We shall cover the two stages in details.

IV. GENERATIVE MODELS

Conversion of text to image is a problem that generative models try to solve. The best model that is being used in this conversion is generative adversarial networks. Below is a brief description of generative models. Consider a dataset of n sampled encoded images as pixel value vector; X= {x1, .xn} where the sampling distribution Pr is unknown, and r represents a real image. A generative model is models that have the ability to learn how to produce samples from the distribution Pg which estimates the sampling distribution Pr. Pg the model distribution is the best guess of the data distribution (the distribution of the data where the sample was sampled from). The generative models maximize the log-likelihood function with respect to θ, ex ~ Pr log (Pg (x| θ) in order to learn the distribution Pg (Ledig, Theis, Huszár, Caballero, Cunningham, Acosta & Shi, 2017). Putting more probability masses around the area of X with fewer samples from the region away from X and more from X is similar to maximum likelihood learning. Minimization of kullback Leibler divergence is similar to the maximization of the log-likelihood function of the model, with the assumption that Pg and Pr are both densities. This method of generative models is preferred since there is no need of knowing the distribution of the population where the sample was sampled from. According to the weak law of large numbers, the expectation will be approximated provided there are enough samples. Most commonly used models are generative adversarial networks. (Zhang, Huang, Wang, &Metaxas, 2017).

V. STACK GENERATIVE ADVERSARIAL NETWORK

This model is used in the conversion of natural language to image conversion. In a modal such as StackGAN, when the model is feed with the image descriptions, in the first stage it produces an image of low resolution by sketching basic color of the described objects and the rough shapes of the objects. After that stage, it enters the second stage where it considers the text description and the output from stage one as its input, then it generates an image that is realistic and of high resolution (Reed, Akata, Yan, Logeswaran, Schiele, Lee, 2016). One of the major challenges this kind of problem faces is the vast number of images that could fit the provided description. In the recent past, the generative adversarial network has improved due to the adjustments made and the results produced from

complex multimodal data synthesis and modeling are promising more than ever. Though, text to image synthesis it produces credible 64by 64 images based on the text description, there are some issues associated with, the images produced usually rich objects parts and image details (Nie, Trullo, Lian, Petitjean, Ruan, Wang, Shen, 2017). In some occasions, they fail to produce vivid bird eyes and beaks. The synthesis also lacks the ability to handle images of high resolutions when additional spatial clarifications of the subject image are not provided.

In solving the problem of natural language to image conversion is decomposed into two main easily convenient sub-divisions using stacked GANs as discussed above where we have stage one and two syntheses. Image of low resolution is produced in the first stage as the model learns to sketch rough drawings applying them with the basic color, in this stage vector of the random noiserandomly selected from the prior distribution is used to generate the background area. The image produced in the first stage usually contains distorted images which have a lot of defects, we shall handle the two stages in details in the adversarial network method section. (Zhang, Huang, Wang, Metaxas, 2017). Then we run stage 2 which outputs high resolution and more realistic images. As mentioned above, the stage 2 of the synthesis makes use of the first stage output and the image text description as its input. (Zhang, Huang, Wang, & Metaxas, 2017). The main work in stage two is to make the images more detailed as possible and remedy the defects. All the information omitted in stage one is utilized in stage two. This makes it easier to draw the image since it's very difficult to draw from scratch a high-resolution image.

The image produced in the first phase (it is distorted and it's just a rough output).

The image produced in the second stage is more realistic and of high resolution (bit depth).

Stacked generative adversarial networks have the capability of producing realistic images from a given image text description in two stages. Comparing this model with other models that were used in the text to image conversion this model outdoes them with almost 2o percent. The stack GAN can be used even in high-resolution image generation, while other models experience challenges when generating an image of 64 by 64 resolution level when additional spatial explanations are not provided.

VI. GENERATIVE ADVERSARIAL NETWORK

The generative adversarial network comes as a countermeasure to solve most of the issues that were experienced with the generative models. Some of the issues included:

- The image quality needed to be stepped up.
- There was a need for a model that never needed the density Pg which in most cases is unknown.
- There was a need for a model that would work in parallel and efficiently.
- Finally, there was need of a model that was flexible based on the network generating samples topology and the loss function.

The convergence equation of Pr and Pg is not biased- does not possess the biasness estimator for the loss it optimizes (Goodfellow, Pouget-Abadie, Mirza, Warde-Farley, Ozair&Bengio, 2014). However, these new models possess two major challenges, the model does not indicate when a convergence occurs and it's unstable especially during training. To maintain the stability of the generative adversarial network there is a need for using a specific architecture and choosing the hyper parameters carefully. This, however, is not ideal, if it will be done in every training process (Zhang, Huang, Wang, &Metaxas, 2017). The generative adversarial framework is usually grounded on a game involving two entities, the generator, and the discriminator/critic. In this game, the generator produces an image and tries to confuse the critic that the image is a real one. Then the discriminator tries to check if the image is real /synthetic (Berthelot, Schumm & Metz, 2017). As the game continues the two learn from experience with respect to the task they are provided with and their performance measure continues to increase with the increased task and hence experience. Eventually, they produce more realistic images (Kim, Cha, Kim, Lee & Kim, 2017). Mathematically, given a dataset X, comprising xi samples that belong to compact matrix x and in space [-1, 1] ^n of the images. The critic learns a function (parametric) DW: x -> [0, 1] which feeds on image xi and produces a probability of how the image is real (Zhang, Huang, Wang, Metaxas, 2017).

Letting J be the range of randomly selected vectors J, which has a fixed; Pz = N (0, 1).

Generator acquires a function Gr: Z -> x that maps the random vector state to the state of the random vector X state (Arjovsky &Bottou, 2017). The images created by the generator corresponds to the states of Pg ~X. hence, the generator is able to acquire knowledge of how to map images to a vector of noise.

In the most occasions, the simplest way of analyzing the game and defining it is that it's a zero-sum game where Gr and Dw are the two players' strategies. Let v (D, G) be the function that represents the discriminator payoff. The generator seeks to minimize v while the discriminator seeks to maximize it. The value of D must be proportionate to that of V in order to extricate real image from the phony image.

$$V(D,G) = E_{X\sim \Pr}[\log(D(x))] + E_{Z\sim Pz}[\log(1 - D(G(z)))]$$

The discriminator tries to push the generator to produce a Pg that is more close to Pr.

VII. GENERATIVE MODELS

To turn the generative adversarial network trivially to a conditional generative model, there must be preset conditions, the vector C must be appended to discriminator and generator in order to generate data (Mirza & Osindero, 2014). The layer to which they are appended does not matter. As the inputs are being added the network will have the capability to learn and adjust its parameters accordingly.

A. TEXT EMBEDDING

For the text to be used in the model – generative adversarial networks- it must first be vectorized. The process of converting text to vector is commonly known as text embeddings. Char CNN-RNN encoder is used in the text embedding computations. The work of the encoder is to map captions and image to a common space of embedding. Images that matches are mapped to highest inner product vectors (Johnson & Zhang, 2015). The convolutional recurrent neural network is concerned with the transformation of the text description while the convolutional neural network is responsible for image processing. Alternatively, a skip Thought vector model can be used which is based purely on language. Sentences with common semantics and syntax are mapped together by this model (Roweis & Saul, 2010). However, char-CNN-RNN is more effective in vision-related tasks since makes use of the matching images from the description. Convolutional features are similar to the embedding of the corresponding images. This aspect makes them visually biased.

B. GAN ARCHITECTURES

Early architectures of GANs consisted of neural networks with fully connected layers; however, they were only good with relatively simple datasets like MNIST and CIFAR-10.A natural next step was to use convolutional neural networks (CNNs) as they are much better suited for images, however, training CNNs for GANs with the same capacity as those used in supervised learning tasks proved to be very difficult. The first major improvement in training GANs with CNNs was DCGAN [6], the authors of DCGAN worked on an extensive search for different. Convolutional architectures and ended up with some guidelines for designing and training generators

GAN Architectures - Early architectures of GANs consisted of neural networks with fully connected layers; however, theywere only good with relatively simple datasets like MNIST and CIFAR-10 (Gulrajani, Ahmed, Arjovsky, Dumoulin, Courville, 2017). A natural next step was to use convolutional neural networks (CNNs) as they are much better suited for images, however training CNNs for GANs with the same capacity as those used in supervised learning tasks proved to be very difficult. The first major improvement in training GANs with CNNs was DCGAN [6], the authors of DCGAN worked on an extensive search for different convolutional architectures and ended up with some guidelines for designing and training generators and discriminators, they use stride and fractionally stride convolutions, which are sometimes called deconvolutions or transposed convolutions, to learn good ways for upsampling and downsampling which eventually results in better images quality (Gulrajani, Ahmed, Arjovsky, Dumoulin, Courville, 2017). They also recommend the use of batch normalization. As it has a stabilization effect while training. Finally, they recommend the use of leaky ReLUs as the activation function, since regular ReLUs usually have sparse gradients which lead to more instability which eventually results in better images quality. They also recommend the use of batch normalization as it has a stabilization effect while training. Finally, they recommend the use of leaky ReLUs as the activation function since regular ReLUs usually have sparse gradients which leads to more instability(Zhang, Huang, Wang, Metaxas, 2017). However, since the focus of this paper is not on models in text embedding, we shall not go into more details.

VIII. RESULTS OF THE RESEARCH

The following section puts forth the results and the findings of the research. It makes available the body of theoretical information regarding the topic of the research, which also provides a detailed review on literature. The images in this research

paper are produced by the generative adversarial model. The codes are generated using python version 3.6 and with the incorporation of the TensorFlow the GPU version. TensorFlow is a python library developed by Google brain researchers and its open source. This library was developed to ensure scientific and numeric computation was done with the highest performance possible. TensorFlow is one of the vastly utilized libraries in machine learning. This is due to the fact that it offers both a high level and low-level APIs making it allow quick iterations and more flexibility (Abadi, Barham, Chen, Chen, Davis, Dean, & Kudlur, 2016). It has the ability of parallel computation by use of modern graphics processing units so as to handle operations effectively and efficiently.

REFERENCES

[1]. Chen, Z., Davis, A., Dean, J., & Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. In Osdi (Vol. 16, pp. 265-283).
[2]. Arjovsky, M., &Bottou, L. (2017).Towards principled methods for training generative adversarial networks. arXiv preprint arXiv:1701.04862.
[3]. Berthelot, D., Schumm, T., & Metz, L. (2017).BEGAN: boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717.
[4]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., &Bengio, Y. (2014).Generative adversarial nets.In Advances in neural information processing systems (pp. 2672-2680).
[5]. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., &Courville, A. C. (2017).Improved training of wassersteingans.In Advances in Neural Information Processing Systems.
[6]. Johnson, R., & Zhang, T. (2015).Semi-supervised convolutional neural networks for text categorization via region embedding.In Advances in neural information processing systems.
[7]. Kim, T., Cha, M., Kim, H., Lee, J. K., & Kim, J. (2017).Learning to discover cross-domain relations with generative adversarial networks. ArXiv preprint arXiv: 1703.05192.
[8]. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., & Shi, W. (2017).Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network.In CVPR.
[9]. Mirza, M., &Osindero, S. (2014).Conditional generative adversarial nets. ArXiv preprint arXiv: 1411.1784.
[10]. Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., &Shen, D. (2017).Medical image synthesis with context-aware generative adversarial networks.In International Conference on Medical Image Computing and Computer-Assisted Intervention.Springer, Cham.
[11]. Radford, A., Metz, L., &Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. ArXiv preprint arXiv: 1511.06434.
[12]. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016).Generative adversarial text to image synthesis. ArXiv preprint arXiv: 1605.05396.
[13]. Roweis, S. T., & Saul, L. K. (2010).Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500), 2323-2326.
[14]. St-Yves, G., &Naselaris, T. (2017).Decoding brain-like representations with a generative adversarial network.In Conference on Cognitive Com-600 putational Neuroscience.
[15]. Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., & Metaxas, D. (2016). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. ArXiv preprint. ArXiv preprint arXiv: 1612.03242, 2(3), 5.
[16]. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. (2017). Stackgan++: Realistic image synthesis with stacked generative adversarial networks. ArXiv preprint arXiv: 1710.10916.
[17]. Denton, E.L., Chintal, S., Fergus, R., Deep generative image models using a laplacian pyramid of adversarial nerworks. In NIPS,2015.
[18]. Radford, A., Merz, L., and chintala, S, Unsupervised representation learning with deep convolutional generative adversarial networks. 2016.
[19]. Mansimov, E., Parisotto, E., Ba, J. L., and Salakhutdinov, R. Generating images from captions with attention. ICLR, 2016.
[20]. Gregor, K., Danihelka, I., Graves, A., Rezende, D., and Wierstra, D.Draw: A recurrent neural network for image generation. In ICML, 2015.

A Bhumgara A Pitale " Text to Image Synthesis in Generative Adversarial Networks" International Journal of Engineering Research and Applications (IJERA), vol. 9, no.1, 2019, pp 09-14