IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134118

Textual Vision Using Quantized Latent Spaces

G. Naga Pavani[1], Mohammed Sahil S[2], Divya Latha K[3], Rajith Bhargav M[4], Karthik Reddy L[5]

Associate Professor Department Of Computer Science and Engineering, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyal, 518501[1]

Computer Science and Engineering, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyal, 518501[2-5]

Abstract: Textual vision, the fusion of natural language processing and computer vision, has gained significant attention in recent years due to its applications in tasks such as image captioning, text-based image retrieval, and visual question answering. In this paper, we explore the utilization of quantized latent spaces in textual vision tasks. Latent space representations, generated from textual data, capture semantic information essential for understanding and interpreting text. By quantizing these latent spaces, we aim to reduce dimensionality while preserving important semantic features. We present a methodology for generating quantized latent space representations from textual data and discuss the process of quantization using various techniques. Experimental results on benchmark datasets demonstrate the effectiveness of our approach compared to baseline methods. Our findings indicate that leveraging quantized latent spaces enhances the performance of textual vision tasks, paving the way for more efficient and interpretable text-based image processing systems.

Keywords: Quantized Latent Spaces, attention mechanism, VQ-VAE mechanism, Conditional GAN, computer vision (CV).

I. INTRODUCTION

Imagine being able to describe an image in words, and then having a computer generate that image for you. This seemingly futuristic concept is at the heart of text-to-image synthesis, a cutting-edge field that holds immense potential for various applications, including content creation, virtual environments, and assistive technologies.

In recent years, one of the most promising approaches to text-to-image synthesis involves the use of generative adversarial networks (GANs), a type of deep learning architecture that pits two neural networks against each other: a generator and a discriminator. The generator learns to create realistic images from textual descriptions, while the discriminator learns to distinguish between real and generated images. Through this adversarial training process, the generator gradually improves its ability to produce high-quality images that align with the given textual input. While GANs have shown remarkable success in text-to-image synthesis, their effectiveness relies heavily on the quality of the latent space representations used to bridge the gap between text and images. Latent spaces serve as a compressed and abstract representation of both textual descriptions and corresponding image features, enabling the generator network to generate visually coherent images from textual inputs.

In this paper, we delve into the realm of text-to-image synthesis using conditional GANs, a variant of GANs that incorporates conditional information, such as textual descriptions, to guide the image generation process. Moreover, we introduce a novel approach that leverages quantized latent spaces to enhance the efficiency and interpretability of text-to-image synthesis models.

Quantized latent spaces offer a compelling solution to the challenges associated with high-dimensional representations by discretizing continuous latent variables into a finite set of discrete levels. This process not only reduces the computational complexity of the model but also enhances interpretability by providing a more structured and interpretable latent space.

Our objective is to investigate the potential of integrating quantized latent spaces into conditional GAN-based text-to-image synthesis frameworks. By combining the expressive power of conditional GANs with the efficiency and interpretability of quantized latent spaces, we aim to push the boundaries of text-to-image synthesis, enabling more seamless and intuitive generation of images from textual descriptions.

DOI: 10.17148/IJARCCE.2024.134118

Throughout this paper, we will present a comprehensive exploration of our proposed approach, including the methodology for generating quantized latent space representations from textual descriptions, the architecture of the conditional GAN model, and empirical evaluations on benchmark datasets.

Our findings aim to shed light on the effectiveness of quantized latent spaces in improving the performance and interpretability of text-to-image synthesis models, paving the way for advancements in this exciting field.

II. LITERATURE REVIEW

[1] The authors of the paper are Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. The title of the paper is "Vector Quantized Diffusion Models for Text-to-Image Synthesis." The VQ-Diffusion model merges VQ-VAE with DDPM for text-to-image synthesis, outperforming AR and GAN methods. It efficiently models discrete latent space, generating detailed and realistic images from text descriptions. Reparameterization enhances computational efficiency in image generation. Advantage: VQ-Diffusion eliminates unidirectional bias and prevents accumulated prediction errors in text-to-image generation. Disadvantage: Limited discussion on potential challenges in scaling the model to larger datasets or more complex scenes.

[2] The authors are Ming Tao and Fei Wu from Nanjing University; Bing-Kun Bao from NJUPT. The title of the paper is “DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis”. DF-GAN employs a one-stage backbone for direct high-resolution image synthesis. It integrates a Target-Aware Discriminator and Deep text-image Fusion Blocks to enhance text-image semantic consistency. The model simplifies the text-to-image process, generating realistic and text-matching images efficiently. The advantage is DF-GAN simplifies the text-to-image process, achieving realistic and text-matching images effectively. The disadvantage is Limited scalability due to computational complexity in high-resolution image synthesis.

[3] The authors are Hiba Imam, Ruba Mutasim, and Ammar Nasr from the University of Khartoum, Sudan. The title of the paper is "SemGAN: Text to Image Synthesis from Text Semantics using Attentional Generative Adversarial Networks". The methodology involves utilizing the Caltech CUB-200 bird dataset for experiments. It includes training a Deep Attentional Multimodal Similarity Model (DAMSM) with LSTM and Inception-v3 CNN for text and image encoding. Generative Adversarial Network (GAN) with an Attention mechanism is employed for text-to-image synthesis and semantic manipulation. The advantage is Enhanced image quality and semantic understanding through whole sentence semantics. The disadvantage is Computational complexity and limited scalability due to dataset size constraints.

[4] The authors are md. zakir hossain 1, (student member, ieee), ferdous sohel 1, (senior member, ieee), mohd fairuz shiratuddin1, hamid laga 1, and mohammed bennamoun 2, (senior member, ieee). The title of the paper is “text to image synthesis for improved image captioning”. The GAN (Generative Adversarial Networks) along with the neural networks such as text encoder and image encoder are used and bi directional LSTM model is used in generating the images along with captions for it. Advantages are that the images generated are highly accurate compared to previous models and they are more synthetically trained.

III. METHODOLOGY

Sequence Of Work

A. *Text Input*

First, users provide detailed descriptions of what they want to see in the image. These descriptions serve as the foundation for generating the image.

B. *Attention Mechanism*

In text-to-image synthesis, attention mechanisms help the model focus on the most important parts of the text or image during the generation process. This means the model can pay attention to specific words or phrases that are crucial for creating the image.

Attention in Text Embeddings: The model uses attention to weigh different words or phrases based on their importance. This helps the model focus on the most relevant parts of the text when generating the image. By doing this, the model can create images that match the input text more accurately.

C. *Quantized Latent Space*

Quantized latent space is a powerful technique used in text-to-image generation tasks. Here's how it works:

DOI: 10.17148/IJARCCE.2024.134118

Latent Space Representation: In traditional methods, the latent space is continuous, but for text to-image tasks, a discrete latent space is more suitable. Quantized latent space divides the space into discrete codes or vectors.

Vector Quantization: This process assigns each point in the continuous space to the nearest code vector. This discrete representation helps capture the categorical nature of textual descriptions.

Text Encoding and Quantization: Textual descriptions are first encoded into a continuous vector space and then quantized by assigning each vector to the nearest code. This discretizes the textual input.

Multimodal Fusion: The quantized latent space acts as a bridge between textual descriptions and images. During generation, the model works in this discrete space, seamlessly combining textual and visual information to create relevant images.

Decoder Network: The decoder network reconstructs images based on the quantized latent codes, capturing the variability and semantics of the textual input.

Training and Optimization: The model is trained end-to-end to minimize the difference between original and reconstructed images using optimization techniques like gradient descent.

D. Conditional GAN

A Conditional GAN (CGAN) modifies the standard GAN architecture by allowing more control over generated images. It achieves this by conditioning the generator with additional information during training, such as labels or specific characteristics. This input helps the generator create images consistent with the conditioning information.

This sequence outlines the steps involved in generating images from textual descriptions using advanced techniques like attention mechanisms, quantized latent spaces, and conditional GANs.

IV. PROPOSED ALGORITHM

Conditional Generative Adversarial Networks (GANs) are like upgraded versions of the usual GAN setup. In traditional GANs, there's a generator that makes fake data from random noise, and a discriminator that learns to tell real data from fake. But with conditional GANs, we add extra information to this process, like labels or other details. This extra info helps the generator create more specific and targeted data.

Introduced in a 2014 paper by Mehdi Mirza and Simon Osindero, conditional GANs have become pretty popular. They've been used for lots of tasks, like turning images into other images, making images from text descriptions, and changing the style of images. These networks are great at generating high-quality data that matches specific characteristics or labels.

How is CGAN Different from GAN?

Conditional Generative Adversarial Networks (CGANs) are a kind of upgrade from the regular Generative Adversarial Networks (GANs). In CGANs, both the generator and the discriminator get extra information during training, like labels.

	GAN	**CGAN**
Generative Model	Generates new samples from a random noise input.	Generates new samples from a random noise input and a class label.
Discriminator	Learns to distinguish between real and generated samples.	Learns to distinguish between real and generated samples and also classify the generated samples based on the class label.

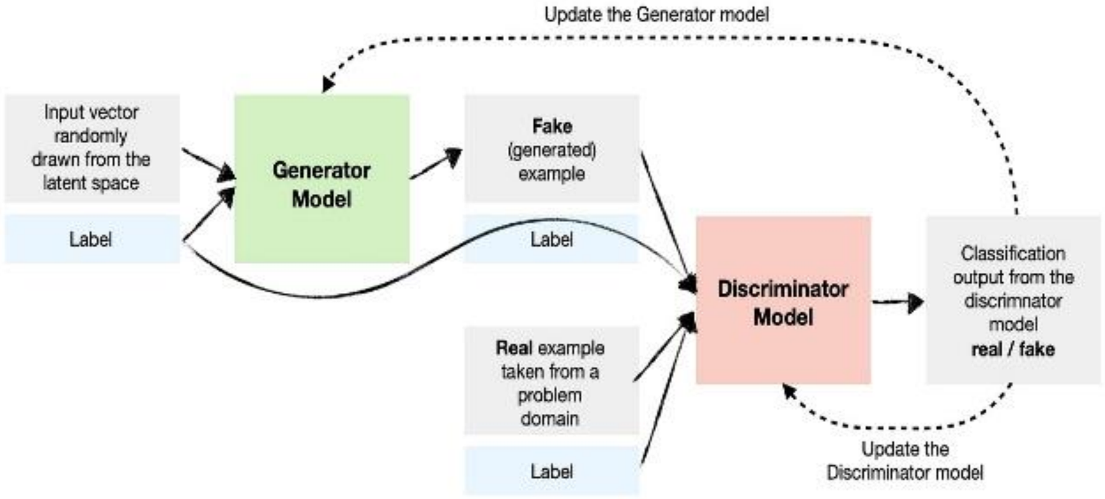
Generator	Learns to generate samples that can fool the discriminator.	Learns to generate samples that can fool the discriminator and also generate samples from specific class labels.
Loss Function	Minimizes the difference between the generated and real samples.	Minimizes the difference between the generated and real samples and also the difference between the generated class label and the input class label.
Applications	Image generation, text generation, etc.	Image generation with class labels, text generation with specific styles, etc.

This extra info helps CGANs make more specific and controlled outputs. Because of this, CGANs can do tasks like changing one type of image into another, making images from text, or altering the style of images. These tasks are tough for regular GANs to handle.

Our methodology involves encoding attention-weighted text through a Vector Quantized Variational Autoencoder (VQ-VAE), introducing discrete codes for efficient representation. Simultaneously, we harness the power of conditional Generative Adversarial Networks (GANs) to refine image generation, optimizing jointly and iteratively. Leveraging the COCO dataset, our system delivers high-quality, diverse outputs closely aligned with user-specified text, showcasing the symbiotic relationship between attention mechanisms, quantized latent spaces, and conditional GANs in enhancing textual vision capabilities for more efficient and interpretable text-based image processing systems.

Architecture of CGAN

The architecture of a Conditional Generative Adversarial Network (CGAN) closely resembles that of a Deep Convolutional Generative Adversarial Network (DCGAN) with some slight modifications. In a CGAN, the Discriminator retains components such as convolutional layers, batch normalization, and Leaky ReLU activation functions from a DCGAN, but it also receives an additional input: conditioning information or labels, alongside the generated image. This additional input empowers the Discriminator to assess both the realism of the image and its adherence to the specified conditions. Similarly, the Generator in a CGAN maintains the framework of a DCGAN but is augmented to accept conditioning information and random noise as inputs. This augmentation enables the Generator to produce images that align with the provided conditions. Overall, the integration of conditioning information enhances the CGAN's capacity to generate images with specific attributes or characteristics.


Update the Generator model
Input vector randomly drawn from the latent space
Label
Generator Model
Fake (generated) example
Label
Discriminator Model
Classification output from the discrimnator model real / fake
Real example taken from a problem domain
Label
Update the Discriminator model

The Discriminator's Network

In a Conditional Generative Adversarial Network (CGAN), the Discriminator's architecture closely resembles that of a standard Deep Convolutional Generative Adversarial Network (DCGAN). It comprises convolutional layers, batch normalization, and Leaky ReLU activation functions. However, a hot-encoding layer is added specifically for the image's labels in a CGAN. This layer encodes conditioning information, such as labels, into a format understandable by the network. By incorporating this layer, the Discriminator evaluates not only the realism of the image but also its alignment with the provided conditions. This capability aids the Generator in producing images with desired characteristics, rendering the CGAN more versatile in generating images with specific attributes.

The Generator's Network

In a Conditional Generative Adversarial Network (CGAN), the Generator's architecture closely resembles that of a Deep Convolutional Generative Adversarial Network (DCGAN). It consists of transposed convolutional layers, batch normalization, and ReLU activation functions. However, an additional layer is included to integrate conditioning information, such as labels. This supplementary layer assists the Generator in generating images aligned with specific characteristics, ensuring greater consistency with the provided conditioning information.

Loss Functions

To train the network, we use two distinct loss functions for the Generator and the Discriminator of the CGAN.

Generator Loss

The objective of the Generator is to gradually improve the quality of fake images by minimizing the disparity between the predicted image and the target. In this architecture, a one-hot encoded label is utilized to determine which features to prioritize. Consequently, the loss function is formulated accordingly.

$$L^{(G)}(\theta^{(G)}, \theta^{(D)}) = -E_z \log D\ (G\ (z\ |y'))$$

Discriminator Loss

The aim of the Discriminator in a Conditional Generative Adversarial Network (CGAN) is to categorize the generated images, with its outputs representing the probability that an image is genuine. The corresponding loss function is formulated as Binary Cross Entropy Loss:

$$L^{(D)}(\theta^{(G)}, \theta^{(D)}) = -\ E_{x\sim Pdata} \log D(x|y) - E_z \log(1-D(G(z|y')))$$

Training

Training a CGAN closely parallels the training process of any other GAN. Both the Discriminator and the Generator operate simultaneously to create new images and determine their authenticity. The Generator begins by generating images from random noise, which are then evaluated by the Discriminator. In a CGAN, the Discriminator is provided not only with images but also with conditioning information, such as labels. This informs its assessment of the generated images' realism. Subsequently, the Generator adjusts its parameters based on the feedback received from the Discriminator, aiming to produce more realistic images. This iterative process continues until the desired image quality is achieved.

Training Flow

The training of a CGAN can be broken down into several steps:

- The Generator initializes with random noise, known as a latent code, as input, which it processes to generate an image.

- The generated image, along with the corresponding conditioning information, is forwarded to the Discriminator. The Discriminator evaluates both the realism of the image and the consistency with the conditioning information, providing a probability score.

- This probability score is then used by the Generator to update its parameters, aiming to minimize the disparity between the generated and real images.

- The process iterates multiple times until the Generator can produce high-quality images. Throughout training, the Discriminator and Generator mutually improve, with the Discriminator enhancing its ability to discern realistic images and the Generator improving its image generation quality.

- It's crucial to note that during training, the Discriminator requires real images and conditioning information to make informed judgments about the generated images' realism and the consistency with the provided conditions.

DOI: 10.17148/IJARCCE.2024.134118

V. EXPERIMENTAL RESULTS

The following are the expected output images for the prompt given by the user.

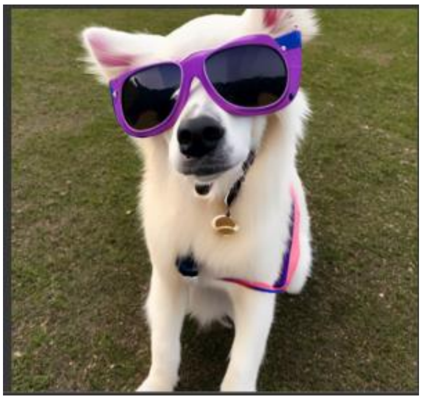


Fig 1. Dog wearing sunglasses

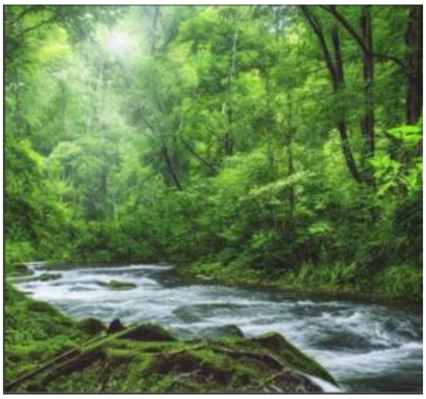


Fig 2. Green forest with river

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose the integration of Conditional Generative Adversarial Networks (CGANs) with attention mechanisms and quantized latent spaces presents a formidable approach in training models for image generation conditioned on textual descriptions.

By harnessing the power of attention mechanisms, the model can effectively capture relevant textual features, while quantized latent spaces enhance efficiency and representation capacity. This blend of cutting-edge techniques from generative modeling and natural language processing promises exciting possibilities. With careful training and fine-tuning, this approach could revolutionize content creation, virtual reality, and interactions with computers, opening doors to new creative horizons.

Moving forward, future advancements in this approach may prioritize the refinement of attention mechanisms and the enhancement of quantized latent spaces to improve the translation of textual descriptions into visually captivating images. Additionally, integrating cutting-edge language understanding models could enhance interpretability and broaden applicability across different domains.

Furthermore, endeavors aimed at enhancing control and interpretability of generated images, alongside exploration into multimodal content synthesis and interactive storytelling, show potential for fostering innovation in diverse fields.

REFERENCES

[1] MD. ZAKIR HOSSAIN 1 , (Student Member, IEEE), FERDOUS SOHEL 1 , (Senior Member, IEEE), M. Z. Hossain et al.: Text to Image Synthesis for Improved Image Captioning. date of publication April 26, 2021, date of current version May 5, 2021.

[2] Ming Tao1 Bing-Kun Bao1,2 Hao Tang3 Changsheng Xu2,4,5 2Peng Cheng Laboratory 3CVL, ETH Zurich ¨ 4University of Chinese Academy of Sciences 5NLPR, Institute of Automation, CAS. GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis. arXiv:2301.12959v1 [cs.CV] 30 Jan 2023

[3] Dominic Rampas, Pablo Pernias, Marc Aubreville. A Novel Sampling Scheme for Text- and Image-Conditional Image Synthesis in Quantized Latent Spaces. arXiv:2211.07292v2 [cs.CV] 23 May 2023

[4] Shuyang Gu1 Dong Chen2 Jianmin Bao2 Fang Wen2 Bo Zhang2 Dong dong Chen3 Lu Yuan3 Baining Guo2. Vector Quantized Diffusion Model for Text-to-Image Synthesis. arXiv:2111.14822v3 [cs.CV] 3 Mar 2022

[5] Ming Tao1 Hao Tang2 Fei Wu1 Xiaoyuan Jing3 Bing-Kun Bao1* Changsheng Xu4,5,6. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. arXiv:2008.05865v4 [cs.CV] 15 Oct 2022.

[6] Ammar Nasr, Ruba Mutasim, Hiba Imam. SemGAN: Text to Image Synthesis from Text Semantics using Attentional Generative Adversarial Networks. 2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE).

DOI: 10.17148/IJARCCE.2024.134118

[7] David Stap∗ Maurits Bleeker Sarah Ibrahimi Maartje ter Hoeve. Conditional Image Generation and Manipulation for User-Specified Content. arXiv:2005.04909v1 [cs.CV] 11 May 2020.

[8] Ming Tao1 Hao Tang2 Fei Wu1 Xiaoyuan Jing3 Bing-Kun Bao1* Changsheng Xu4,5,6. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. arXiv:2008.05865v4 [cs.CV] 15 Oct 2022.

[9] Tingting Qiao1,3, Jing Zhang2,3,*, Duanqing Xu1,*, and Dacheng Tao3. MirrorGAN: Learning Text-to-image Generation by Redescription. arXiv:1903.05854v1 [cs.CL] 14 Mar 2019.

[10] Zixu Wang1 , Zhe Quan∗1 , Zhi-Jie Wang∗23, Xinjian Hu1 and Yangyang Chen1. TEXT TO IMAGE SYNTHESIS WITH BIDIRECTIONAL GENERATIVE ADVERSARIAL NETWORK. 978-1-7281-1331-9/20/$31.00 c 2020 IEEE.

[11] Li Xiaolin1,a,b,c, Gao Yuwei*2,a,b. Research on Text to Image Based on Generative Adversarial Network. 2020 2nd International Conference on Information Technology and Computer Application (ITCA).

[12] Tingting Qiao1,3, Jing Zhang2,3,*, Duanqing Xu1,*, and Dacheng Tao3. MirrorGAN: Learning Text-to-image Generation by Redescription. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).