Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Advances in Text-to-Image Generation: Integrating Transformer Models and Self-Attention Mechanisms

Chang Yuan *

College of Computer Science and Technology, Harbin Engineering University, Harbin, China

* Corresponding author: yc18705277753@hrbeu.edu.cn

Abstract. This study provides a comprehensive overview of advancements in Text-to-Image (TTI) generation through the application of Transformer models and Self-Attention Mechanisms. It begins with a review of the evolution of Generative Adversarial Networks (GANs) and highlights the benefits introduced by Self-Attention, such as improved contextual understanding and clearer image generation. The paper explores the theoretical foundations of Transformers and GANs, detailing how their integration can enhance TTI tasks. It also examines several leading models that employ these methodologies and presents quantitative performance evaluations comparing these models with other commonly used approaches. The findings indicate that Transformer-based modifications significantly improve TTI performance. The study concludes by assessing the current state of Self-Attention techniques and identifying potential research directions, such as exploring multi-head, hard, and soft attention mechanisms. These future research efforts are expected to further refine TTI capabilities and address existing challenges, providing deeper insights and more robust solutions for generating diverse and high-quality images from textual descriptions.

Keywords: Text-to-Image; Transformer Models; Self-Attention Mechanism; GANs.

1. Introduction

Computer vision is one of the most typical application scenarios of neural network models. Interdisciplinary tasks are raised combining Natural Language Processing and computer vision, including text-to-image (TTI) [1], Visual Question Answering (VQA), Image Caption, and Visual Language Navigation. TTI is a typical Multimodal Machine Learning task. With appropriately trained deep learning models, it is capable of generating clear, accurate images according to the description provided. It is anticipated to perform well in applications such as artistic design and generating visual features of criminals. Meanwhile, the progress of improving TTI quality also promotes the development of natural language processing and deep learning models.

So far, researchers have introduced deep learning networks to complete the TTI task. The most widely applied framework is generative adversarial network (GAN) [2]. GAN is a model constructed based on neural networks. It is the mainstream of solving Multimodal Machine Learning tasks. Researchers introduced the convolutional neural network (CNN) and recurrent neural network (RAN) to capture the semantic relationships of the text description [3,4]. CNN uses gradient-based learning algorithms to process the text input. In the early 21st century, CNN is recognized as the future of computer vision. However, when it comes to natural language processing, CNN performs much less than satisfaction. Due to its multilayer structure, each feature a layer capture is affected by the previous one, which is a lineal calculating procedure. CNN is also highly relevant to iterative algorithms. These features determine that frameworks using CNN are incapable of training parallelly and capturing semantic relationships from a global perspective. In 2018, Nick Seaver published the pioneering essay Attention [5]. It introduced the Self-attention Mechanism to substitute for the multi-layer framework and opened up a new era for the computer vision field.

In 2018, researchers introduced the Attention Mechanism to the GAN framework, which gave birth to Attentional Generative Adversarial Networks (AttGAN) [6]. In the same year, Google released the Image Transformer, which marks the former application of the Transformer in the computer vision field [7]. Finally, in 2020, Researchers combined Self-attention Mechanism and AttGAN, and

established the generative adversarial network based on Transformer [8]. The model has become the mainstream for TTI tasks and has been widely applied.

The primary objective of this study is to introduce the development and core theories of TTI generation based on Transformer models and to highlight the innovations in this area. The first section of the article reviews the evolution of GANs and explains the advantages brought by the incorporation of the Self-Attention Mechanism. The discussion then shifts to the theoretical foundations of Transformers and GANs, detailing how these frameworks can be combined to enhance TTI tasks. Subsequently, the article presents several prominent models that utilize these methods. Performance evaluations are provided, comparing these networks with other widely used models to quantitatively demonstrate their advantages. In the final section, the article assesses the current state of Self-Attention techniques and outlines future research directions in this field.

2. Methodology

2.1. Dataset Description and Preprocessing

Currently, researchers employ multiple data sets to assist in training and evaluating deep learning models of TTI. The most commonly used data sets are MS COCO and CUB [9,10]. Microsoft Common Objects in Context (MS COCO) is a large-scale data set that provides images of various objects with the corresponding description. Caltech-UCSD Birds-200-2011 (CUB) is a data set specialized in birds. It provides images and descriptive text of 200 species of birds, mainly applied in fine-grained image research. Additionally, Text2Scene is widely used when scenario materials are required [11].

2.2. Proposed Approach

Text-to-image is a natural language processing task applying deep learning models. Generative adversarial network functions via a series of neural networks. However, most of the mainstream solutions mainly employ CNN and RNN, which contain inevitable flaws such as the occurrence of Vanishing Gradient when handling long-range text sequences [12]. Thus, it is a significant innovation to introduce Transformer. The Self-attention Mechanism has brought massive changes to NLP tasks and extensively improved the results of TTI. After inputting text descriptions, the system turns texts into processable forms by methods including Word Embedding [13]. Subsequently, generating models such as GAN are used to generate corresponding images according to the input. Finally, the generated images are generated and refined. The pipeline goes as shown in Fig. 1.

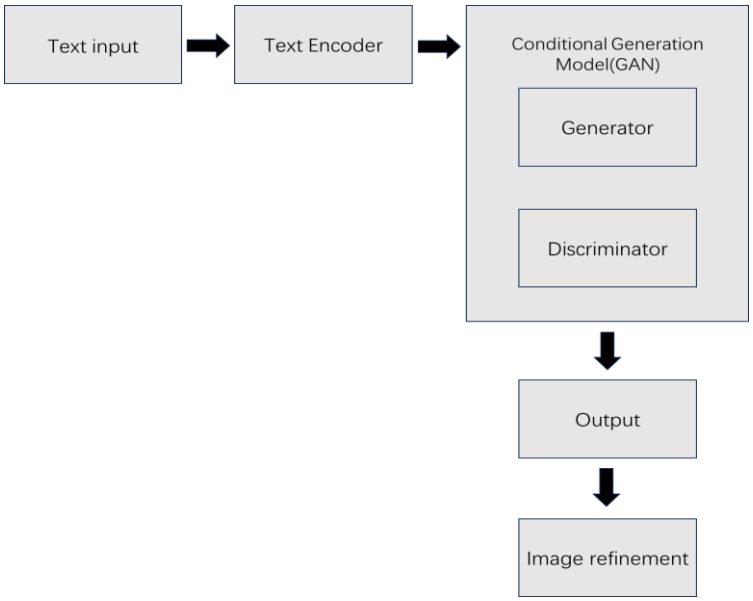

Figure 1. the pipeline of TTI

2.2.1. Basic theories of GAN.

TTI tasks require models able to understand and obtain the main idea of the text. Thus, GAN is the most important part of it since GAN is responsible for analyzing inputs and generating images. Generative adversarial network is established based on deep learning models. It mainly consists of two kinds of neural networks, generator and discriminator. The generator consistently generates fake images via the semantic relationships while the discriminator compares the fake images with its pre-trained realistic visual data set. The discriminator evaluates the fake images and determines whether it is fake or not. When the generation and evaluation reach Nash Equilibrium, it means both the generator and discriminator have achieved the optimal strategy [14]. Traditional GAN employs step-by-step region-level modification to maintain visual consistency across long-range sequences [15]. Each layer of the convolutional neuro network highly relies on the previous information. After several times of convolution, the initial information may get lost. A similar situation occurs when RNN is applied because of the reliance on time series [16].

2.2.2. Transformer modification.

Transformation is an innovational deep learning neural network according to Attention [5]. Transformer is composed of inputs, outputs, an encoder, and a decoder. The model is shown in Fig. 2.

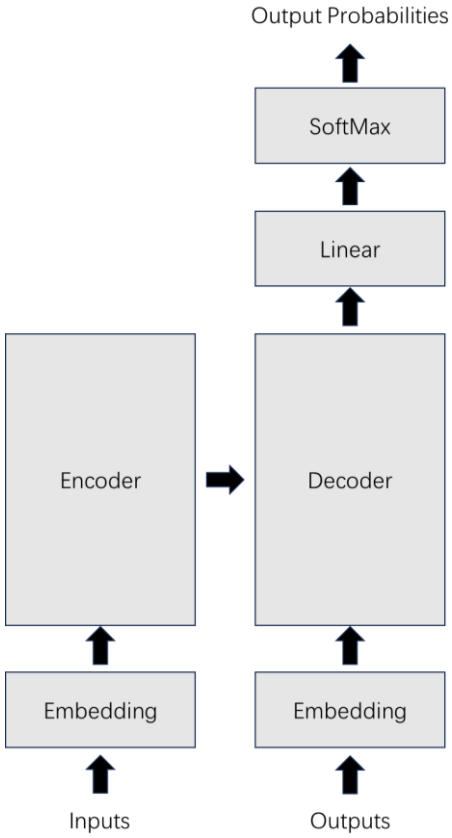

Figure 2. Transformer modification

The encoder-decoder modification has substituted for the iterative algorithm, which is the key of Transformer. Both Encoder and Decoder are divided into six parallel layers, enabling the inputs to be processed simultaneously. After receiving the inputs, the embedding algorithm captures the positional encoding of the text and adds it to the word embedding. The sum continuously performs as the computational object. Afterward, the word embedding matrix is sent into the encoder-decoder and processed by the Muti-head Attention layer. Finally, the feed-forward layer and SoftMax function emit the outcome.

Position encoding and Self-Attention Mechanism are the most important features of Transformer. Position encoding enables the model to capture semantic relationships from a global perspective. In

the traditional gradient-based algorithm, the inputs are recognized as chronological sequences. The text is processed according to the lineal order consequently.

2.2.3. Self-Attention Mechanism (GAN-SelfAtt).

To further improve the quality of the generated images, researchers introduced the generative adversarial network based on GAN-SelfAtt [17]. The generating network includes a generator based on the deconvolutional neural network; a discriminator based on the convolutional neural network and a self-attention layer. The generator and discriminator operate similarly to the ones of traditional GAN modifications. Additionally, Huang introduced the self-attention module as a substitute for the convolutional layers. The feature map generated from the formerly hidden layer is input into the feature space to calculate the attention value. In the self-attention module, each area that receives attention is processed individually. The results are then computed for the weighted sum. The weighted sum is ultimately added up with the original feature map as the final result. In this way, the model is capable of diminishing the flaws of convolutional network and performing better in long-ranged tasks.

2.2.4. Cross Attention Encoder Generating Adversarial Network (CAE-GAN).

CAE-GAN is a hybrid model raised in 2022. CAE-GAN has applied several deep-learning networks together. It is a Multimodal model capable of handling multiple NLP tasks. The main structure of the model includes a Pre-trained Encoder, Dynamic Memory Module, and Three-level GAN. The Pre-trained Encoder employs the Cross-Attention Mechanism [18]. Researchers designed the cross-attention encoder to deal with the semantic relationships of text description and image generation. The encoder is separated into four modules to individually capture text features, generate image features, and encode according to the Cross-Attention Mechanism and Self-Attention Mechanism. The encoder goes as shown in Fig. 3.

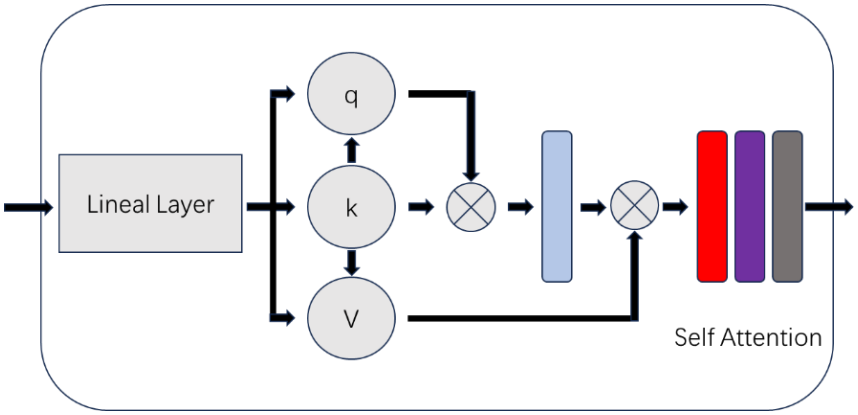

Figure 3. Cross-Attention Encoder

The Self-Attention layer replaces the middle part of the convolutional layers. By utilizing the Transformer, it is possible to get quick results to improve its performance of capturing semantic relationships and bring remarkable changes to the image quality.

3. Result and Discussion

This chapter reveals the quantitative comparison between CAE-GAN and the mainstream generating models. According to the result, it is clear to distinguish whether introducing Transformer holds superiority. The chapter continues to discuss the further development of TTI based on Transformer and its future outlook.

3.1. Quantitative Comparison

3.1.1. Experiment.

Researchers applied two subtests to go on the experiment. The datasets CUB and COCO were employed to train and test the models. The details of the datasets are shown in Table 1.

Table 1. Dataset details

Dataset	Trainset	Testset
Coco	8855	2933
CUB	82783	40470

The experiment involved three steps, pre-training, comprehensive training, and testing. The pre-training of cross attention encoder was the main focus of the procedure. The encoder is required to be trained individually according to each subtest. It can enable the model to work more flexibly directly against specific datasets to capture the featured mapping relationships between images and texts. The trained encoder had been loaded in the first place when training comprehensively. Then the rest parts were trained separately. In the testing phase, researchers generated 30,000 images based on CUB and COCO. The Inception Score (IS) and Fréchet Inception Distance (FID) were calculated accordingly to evaluate the performance of the model quantitatively.

3.1.2. Evaluating index.

Researchers employed the IS and FID to evaluate the performance of the model [19]. IS is an important reference in GAN evaluation. Higher IS represents higher image quality and broader variety. FID reveals the distance between real images and the generated images. Lower FID shows that the generated images hold closer features to the real ones, which means the model is capable of generating more realistic pictures.

3.1.3. Experiment results analysis.

Researchers used CAE-GAN to generate 30,000 pictures according to CUB and COCO. The results are compared with mainstream models such as Stacked GAN (StackGAN), AttGAN, and Dynamic Memory GAN (DM-GAN). The results are shown in Table 2 and Table 3.

Table 2. Results using CUB

Models	IS	FID
StackGAN	3.70±0.04	35.11
AttGAN	4.36±0.03	23.98
DM-GAN	4.75±0.07	16.09
SegAttnGAN	4.82±0.05	-
CAE-GAN	4.87±0.06	13.66

Table 3. Results using coco

Models	IS	FID
StackGAN	8.45±0.03	-
AttGAN	25.83±0.47	35.49
DM-GAN	30.49±0.57	32.64
objGAN	30.29±0.33	-
OP-GAN	28.57±0.17	-
CAE-GAN	30.96±0.56	30.83

Table 2 shows the results of CAE-GAN with other mainstream models while tested according to CUB. Compared to the traditional DM-GAN, the IS of CAE-GAN rises from around 4.75 to around 4.87. The score also improved by 1.04% according to Sequence GAN (SegAttnGAN). The CAE model presents better visual quality. The FID of CAE-GAN is 13.66, which is lower than DM-GAN and AttGAN. It is revealed that CAE-GAN generates more realistic pictures. Table 3 shows the IS and FID of StackGAN, AttGAN, DM-GAN, Object-driven Attentive GAN (objGAN), Open Set GAN (OP-GAN), and CAE-GAN tested with COCO. The IS of CAE-GAN is about 30.96, while the FID

is 30.83. Both the scores perform better than the traditional models. Using the quantitative experiment, it is easy to see that CAE-GAN achieved an impressive improvement in image quality and variety compared to the traditional models. The clarity and realism both reach a higher standard. Hence, introducing the Self-Attention Mechanism into cross attention network performs better in TTI tasks.

3.2. Discussion

Compared to traditional GANs, models that incorporate the Self-Attention Mechanism offer a more nuanced understanding of context. This leads to the generation of images with greater clarity and more accurate features. The Transformer architecture has seen widespread application in TTI tasks due to its ability to capture long-range dependencies and intricate relationships within data. However, a significant challenge that researchers currently face is enhancing the diversity of the generated images.

Despite the advances made, the issue of image variety remains prominent. While Self-Attention Mechanisms improve the fidelity and contextual relevance of generated images, ensuring a broad spectrum of visual outputs from diverse textual inputs is still complex. Researchers must address how to expand the range of generated images while maintaining high-quality results. This involves exploring new techniques for improving model robustness and versatility, such as incorporating more diverse training datasets, refining network architectures, and optimizing loss functions to promote variability. Continued innovation in these areas is crucial for advancing the capabilities of TTI models and achieving more diverse and realistic image generation.

4. Conclusion

This study explores the application of Transformer and GAN theories in the context of TTI tasks, with a particular focus on the Self-Attention Mechanism and its advantages. The aim is to provide readers with a comprehensive understanding of how Self-Attention enhances TTI models, thereby encouraging further exploration in this area. The study highlights two representative GAN models that utilize Self-Attention and presents a quantitative evaluation of the CAE-GAN model. The findings demonstrate that modifications based on Transformers can significantly improve TTI tasks. Looking ahead, future research will likely focus on other attention mechanisms, such as multi-head attention, hard attention, and soft attention. The next stage of research will aim to examine the unique features and applications of these various attention mechanisms, exploring how they can further advance the capabilities of TTI models and address current limitations. This continued investigation will contribute to a deeper understanding of attention-based architectures and their potential to enhance image generation from textual descriptions.

References

[1] Reed S. Akata Z. Yan X. et al. Generative adversarial text to image synthesis. International conference on machine learning. PMLR, 2016: 1060 - 1069.

[2] Goodfellow I. Pouget-Abadie J. Mirza M. et al. Generative adversarial nets. Advances in neural information processing systems, 2014, 27.

[3] Le C.Y. Bottou L. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86 (11): 2278 - 2324.

[4] Tsai Y.H. Bai S. Liang P. Multimodal transformer for unaligned multimodal language sequences. 2019, arXiv print: 1906. 00295.

[5] Vaswani A. Shazeer N. Parmar N. et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017: 5998 - 6008.

[6] Xu T. Zhang P. Huang Q. et al. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[7] Dosovitskiy A. Beyer L. Kolesnikov A. et al. An image is worth 16×16 words: transformers for image recognition at scale. 2020, arXiv print: 2010. 11929.

[8] Tan X.Y. He X.H. Wang Z.Y. et al. Text generation image technology based on Transformer cross-attention. Computer science, 2022, 49 (02): 107 - 115.

[9] Lin T.Y. Maire M. Belongie S. et al. Microsoft coco: Common objects in context. Computer Vision–ECCV European Conference, 2014: 740 - 755.

[10] Wah C. Branson S. Welinder P. et al, “The caltech-ucsd birds-200-2011 dataset”, 2011, http://www.birdfifieldguide.com.

[11] Tan F. Feng S. Ordonez V. Text2scene: Generating compositional scenes from textual descriptions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 6710 - 6719.

[12] Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 1998, 6 (02): 107 - 116.

[13] Kusner M. Sun Y. Kolkin N. et al. From word embeddings to document distances. International conference on machine learning. PMLR, 2015: 957 - 966.

[14] Daskalakis C. Goldberg P.W. Papadimitriou C.H. The complexity of computing a Nash equilibrium. Communications of the ACM, 2009, 52 (2): 89 - 97.

[15] Cheng Y. Gan Z. Li Y. et al. Sequential attention GAN for interactive image editing. Proceedings of the 28th ACM international conference on multimedia. 2020: 4383 - 4391.

[16] Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 1998, 6 (02): 107 - 116.

[17] Huang H.Y. Gu Z.F. A text image generation adversarial network based on self-attention mechanism. Journal of Chongqing University, 2020, 43 (03): 55 - 61.

[18] Hou R. Chang H. Ma B. et al. Cross attention network for few-shot classification. Advances in neural information processing systems, 2019, 32.

[19] Barratt S. Sharma R. A note on the inception score. 2018, arXiv preprint: 1801. 01973.