Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 9, No. 3, 2024

Research on Handwriting Text Generation Algorithm Based on Generative Adversarial Network

Fuchang Zhao

School of Electronic and Information Engineering, Southwest Minzu University, Chengdu, 610229, China

Abstract: The main task of this paper is to study the handwriting text generation method based on deep learning. Through understanding the development status of the research, It can be found that the current research on the generation of different handwriting styles still has some obvious defects, such as the need for manual intervention in character segmentation, failure to capture the global handwriting style, style collapse, failure to generate arbitrary length characters, and text. Finally, this paper proposes a handwritten text generation algorithm combining advantages of convolutional network and Transformer. Specifically, this paper first constructs a lightweight backbone network and uses lightweight MobileNetv3 network as the backbone network to realize feature extraction of input images. The Efficient Channel Attention module is introduced to replace the SE attention module of MobileNetv3, which makes the network pay more attention to the global and local style features of handwritten images. In the feature extraction part, the network reduces the number of parameters, calculation amount and video memory occupation. It can also extract rich feature information. The FID and KID indexes of this algorithm obtained 20.28 and 9.07×10^{-3} respectively, and the generation effect of handwritten pictures was excellent, which could effectively imitate the writing style of writers.

Keywords: Deep learning, GAN, handwritten text generation, multimodal, picture generation.

1. Introduction

Writing is an extraordinary achievement of human beings, and the generation and development of writing represents a milestone of human civilization. For a long time, beautiful handwriting and calligraphy have not only been regarded as an art form of language, but also the use of deep learning methods to generate handwritten texts of specific authors has both historical significance and wide application prospects. Practical applications of this research range from synthesizing high-quality training data for the training of personalized handwritten text recognition[1,2,3,4] models to automatically generating handwritten notes for people with physical impairments, providing more accessible cultural experiences.

Stylized handwritten text generation (HTG) is an emerging research field, a comprehensive task that combines the fields of computer vision (CV) and natural language processing (NLP). Designed to generate specific handwritten text images that mimic the author's calligraphic style. The challenge of handwriting style generation research is that the algorithm needs to understand and capture the author's unique writing style, so as to generate similar handwriting text, and simply adopting style transfer is not feasible. In fact, calligraphy that mimics a particular author involves not only texture, nor the thickness, tilt, skew, and roundness of strokes, but also the shape and ligature of individual characters. In addition, these visual aspects must be handled correctly to avoid artifacts that can cause content changes (for example, even small additional or missing strokes).

The realm of handwritten text generation has garnered substantial research attention[5,6,7,8,9]. Alonso et al. [10]proposed an approach employing a bidirectional LSTM loop layer that encodes the character sequence for generation, using the input content string as a condition. To control text content in generated images, they introduced an auxiliary network for text recognition. However, their method is confined to training on isolated fixed-size word images, posing challenges in generating high-quality arbitrary length text.

Recently, GAN-based techniques have been introduced to address the issue of stylized handwritten text image generation. These methods consider both content and style [11,12]when generating offline handwritten text images. Davis et al. presented a StyleGAN-based approach that learns to generate handwritten image widths based on style and input text. Conversely, the GANwriting framework conditions the handwritten text generation process on text content and stylistic features in scenarios with few samples. Nevertheless, both methods face crucial challenges impacting the generation of stylistic handwritten text images.Firstly, the loose relationship between style and content in these methods, treating their representational characteristics separately and linking them later, poses an issue. Although this scheme enables entanglement between style and content at the word/line level, it lacks explicit enforcement of style-content entanglement at the character level. Secondly, while these techniques capture global writing styles like ink width and writing tilt, they do not explicitly encode local style patterns such as character styles and hyphens. Consequently, accurately mimicking native calligraphy style patterns in the reference style examples becomes challenging.

Another approach is ScrabbleGAN[13], which uses a fully convolutional architecture to synthesize handwritten words, achieving impressive results in terms of content, but neither approach is adapted to a particular author's handwriting style.In addition, another approach is to use the Transformer model, proposed by Kumar et al., the first generation network to introduce Transformer for stylized handwritten text generation. However, the use of traditional Transformer also brings a huge increase in the number of parameters and computation.

1.1. Contribution

This paper introduces a generative adversarial network (GAN) based on the Transformer design, enhancing the backbone network of the generator by employing the

lightweight MobileNetV3 backbone and integrating the Efficient Channel Attention (ECA) mechanism to improve the network's focus on handwritten text targets. The generator network in this paper adopts an encoder-decoder structure. After obtaining feature maps through a convolutional network from the input handwritten text images, the feature maps undergo flattening, reshaping, and concatenation operations. Subsequently, the Transformer encoder network utilizes a multi-head self-attention mechanism to generate a self-attentive style feature sequence of the author. This feature sequence is then input into the decoder network, which consists of multi-head self-attention and encoder-decoder attention, to generate character-specific style attributes based on a set of query word strings. Finally, the output obtained from the Transformer decoder is fed into a fully convolutional decoder to generate the final stylized handwritten text image. Additionally, we enhance the style consistency of the generated text by constraining the decoder output through a loss function, aiming to re-generate the author's style feature sequence at the encoder. The proposed GAN model in this paper simulates the author's style given a query content through multi-head self-attention and encoder-decoder attention mechanisms, emphasizing relevant self-attentive style features with respect to each character in the query. This enables the capture of style-content entanglement at the character level. Furthermore, the self-attentive style feature sequence generated by our encoder encompasses both the global (e.g., ink width, slant) and local styles (e.g., character style, ligatures) of the author's handwritten text within the feature sequence.

2. Related Work

Handwritten text can be seen as a trace that captures the stroke shapes that make up a character, or a still image that captures its overall appearance. According to this concept, the research of handwritten text generation can be divided into online handwritten text generation and offline handwritten text image generation.

The online HTG[14,15] method utilizes a sequential model, such as LSTMs, Conditional Variational RNNs, or Stochastic Temporal CNNs, to predict the position of the pen point by point based on the current position and the input text to be rendered. The earliest methods to adopt this strategy have no control over style. This limitation was then addressed by subsequent work to decouple and then recombine content and author styles. The main disadvantages of online HTG methods are that they are difficult to learn long-term dependencies, and they require training data composed of digital transcripts, which are difficult to collect and may not even be available in the application scenario of historical manuscripts. Therefore, in this paper, we employ off-line HTG methods, which typically allow inference of stylistic glyphs that are not directly observed in style examples by using GANs for either non-styled or conditional HTG.

Learning-based solutions rely on Gans, either unconditionally (for non-styled HTG) or conditional on a different number of handwritten style samples (for styled HTG). In the latter case, the style sample can be an entire paragraph or line, a few words, or a single word. The first method is able to generate fixed-size images based on content embedding, but has no control over the calligraphy style. Unlike natural image generation, generating handwritten text images requires generating variable-size images. Therefore, the method proposed in ScrabbleGAN aims to overcome this limitation by linking character images, but still cannot mimic handwriting style.

Unlike natural image generation, generating handwritten text images should involve generating variable-size images. Thus, the method proposed in ScrabbleGAN aims to overcome this limitation by concatenating character images, but is still unable to mimic calligraphic styles. The solution to style HTG is not only generated on text content, but also on vector representations of styles. In the HIGAN,and HIGAN+ methods[16,17], the stylistic representation and content representation of text images are obtained respectively, and then combined in the later stage for generation. This prevents these methods from capturing native writing styles effectively. The Transformer-based approach proposed by the HWT approach addresses this limitation, better capturing content-style entanglement by leveraging a cross-attention mechanism between style representation and text representation of content.

The handwriting generation countermeasure network in this paper is also mainly based on Transformer, and the backbone network and decoder network are redesigned, aiming to make the model better adapt to the feature extraction and fusion of handwritten text, so as to improve the generation quality.

3. Proposed Approach

This paper proposes a Generative Adversarial Network (GAN) based on Transformer design, enhancing the backbone network of the generator. It improves focus on handwritten text targets by employing a lightweight MobilenetV3 backbone network and introducing the Coordinate Attention mechanism. The generator network in this paper follows an encoder-decoder architecture. After obtaining feature maps through a convolutional network from the input text images, operations such as flattening, reshaping, and concatenation are applied to the feature maps. Subsequently, the Transformer encoder network utilizes a multi-head self-attention mechanism to generate the author's self-attention style feature sequence. This feature sequence is then input into the decoder network, consisting of multi-head self-attention and encoder-decoder attention, to generate character-specific style attributes for a given set of query word strings. Finally, the output from the Transformer decoder is fed into a fully convolutional decoder to generate the final styled handwritten text image.

3.1. Backbone network

In order to avoid the video memory consumption and excessive computing load of traditional deep neural networks, we propose a lightweight convolutional backbone network design that can also take advantage of the expression capability of Transformer in CNN feature space. This design takes into account the small sample Settings and the need for large amounts of data in handwriting generation tasks, avoiding the infeasibility and quadratic complexity of directly applying Transformer or VIT-based designs.

MobileNetV3 is a deep neural network model introduced by Google in 2019, and MobileNetV3 performs well in tasks such as mobile image classification, object detection, and semantic segmentation. The model uses a series of innovative technologies, such as the Squeest-and-Excitation （SE） module for the channel's attention mechanism and the Neural

Architecture Search (NAS) approach, which further improve the performance of the network.figure 1 shows the SE network structure

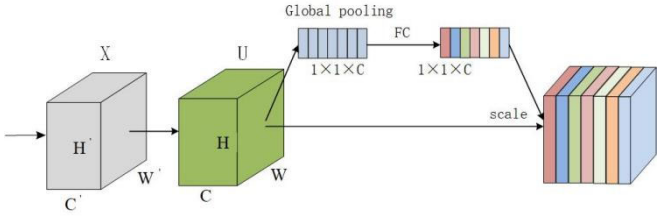

Figure 1. Squeest-and-Excitation module

Figure 1 shows the structure of bneck, the basic module of MobilNetV3. The inverted residual block structure redesigned by MobileNetv3 network adds SE attention mechanism between deep convolution and point convolution on the basis of MobileNetv2, and both use h-swish function as activation function instead of the original ReLU.

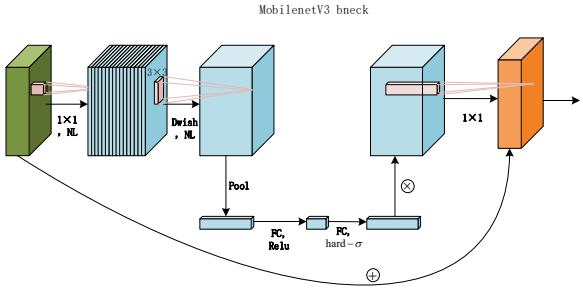

Figure 2. MobileNetV3 bneck

Figure 2 shows the structure of bneck, the basic module of MobilNetV3. The inverted residual block structure redesigned by MobileNetv3 network adds SE attention mechanism between deep convolution and point convolution on the basis of MobileNetv2, and both use h-swish function as activation function instead of the original ReLU.

The results show that the dimensionality reduction operation adopted by SENet will have a negative impact on the prediction of channel attention, and the acquisition of dependency relationships is inefficient and unnecessary. Based on this, an efficient channel attention (ECA) module for CNN is proposed, which avoids dimension reduction and realizes cross-channel interaction effectively.

3.2. Generator

For the construction of generator network, we first identified two important features to consider when designing a method to generate handwritten text of different lengths and any desired style in a small sample context, without using character-level annotations, which further promoted our proposed generator network approach based on Transformer. The first feature is style-content entanglement, and the second feature is global and local style imitation. Transformer can effectively solve these two problems by using multi-head self-attention mechanism and encoder-decoder multi-head attention mechanism. Figure 3 shows the Transformer network structure.

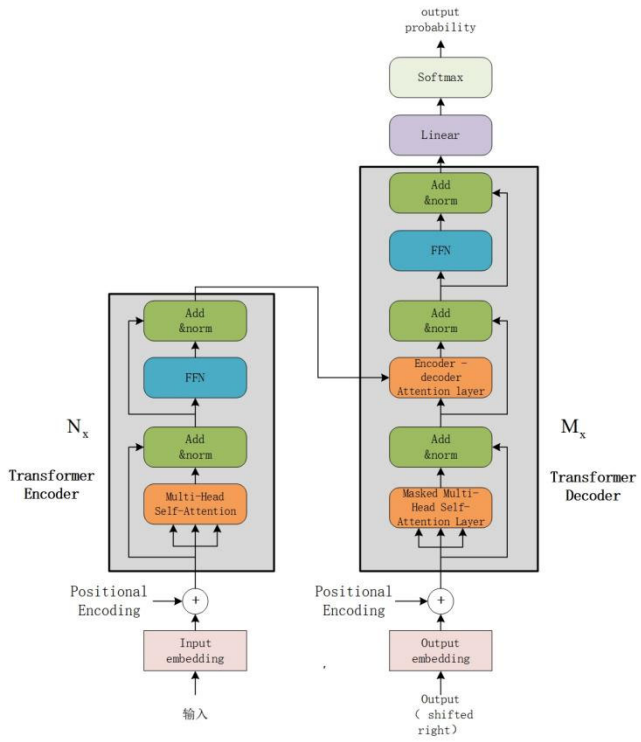

Figure 2. Transformer network architecture diagram

Transformer outputs feature sequences, reshape them, and then generates a complete handwritten text image through convolutional networks and upsampling.

4. Experiments

The experiments were conducted on the IAM dataset with a training epoch set to 4000 and a batch size of 8. The initial learning rate was set to 0.00005, and both the generator and discriminator utilized the Adam optimization method to optimize network parameters. In the Adam optimization algorithm, the coefficients for the momentum term (beta1) and RMSProp term (beta2) were set to 0.0 and 0.999, respectively. The weight decay parameter was set to 0.

Table 1. Comparative experiment of backbone network

Backbone	FID	KID	IS
VGG16	23.72	9.78×10^{-3}	2.79
MobileNetV3	22.59	9.10×10^{-3}	2.56
MobileNetV3+ECA	**20.28**	**9.07×10^{-3}**	**2.53**

In the IAM training of data set, comparative experiments were conducted on models using different convolutional backbone networks, and it can be seen that the FID index IS and KID index of the MobilNetV3+ECA backbone network were improved. Compared with the VGG16 network, the FID value reduced by 3.44, and the indexes of KID and IS are also reduced by 0.71 and 0.26 .Compared with MobilNetV3, which has not been improved, the FID index has also decreased to 2.31, and the KID and IS have also decreased slightly. respectively. Therefore, it can be shown that the improved MobilNetV3+ECA backbone network is adopted in this paper to capture the feature information of images more effectively, so as to produce higher quality generated images in the task of generating adversarial networks.

5. Conclusion

In this paper, the handwriting generation network based on Transformer mainly redesigns the backbone network and decoder network, aiming to make the model better adapt to the feature extraction and fusion of handwritten text, so as to improve the generation quality. In this paper, a lightweight backbone network is constructed, which uses the lightweight MobileNetv3 network to achieve feature extraction of input images, and ECA module is introduced into it, mainly to make the network pay more attention to the global and local style features of handwritten images. In the feature extraction part of the network, on the basis of reducing the number of parameters, calculation amount and video memory consumption, It can also extract rich feature information.

References

[1] Ayan Kumar Bhunia, Abhirup Das, Ankan Kumar Bhunia, Perla Sai Raj Kishore, and Partha Pratim Roy. Handwriting Recognition in Low-Resource Scripts Using Adversarial Learning. In CVPR. IEEE, 2019. 1

[2] Ayan Kumar Bhunia, Shuvozit Ghose, Amandeep Kumar, Pinaki Nath Chowdhury, Aneeshan Sain, and Yi-ZheSong. MetaHTR: Towards Writer-Adaptive Handwritten Text Recognition. In CVPR, 2021. 1

[3] Yue Jiang, Zhouhui Lian, Yingmin Tang, and Jianguo Xiao. SCFont: Structure-Guided Chinese Font Generation via Deep Stacked Networks. In AAAI, 2019. 2

[4] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Mubarak Shah. Handwriting Transformers. In ICCV, 2021. 1, 2, 3, 4, 6, 7, 8

[5] Emre Aksan and Otmar Hilliges. STCN: Stochastic Temporal Convolutional Networks. In ICLR, 2018. 1, 2

[6] Emre Aksan, Fabrizio Pece, and Otmar Hilliges. DeepWriting: Making digital ink editable via deep generative modeling. In CHI. ACM, 2018. 1, 2

[7] Alex Graves. Generating Sequences with Recurrent Neural Networks. arXiv preprint arXiv:1308.0850, 2013. 1, 2

[8] Bo Ji and Tianyi Chen. Generative Adversarial Network for Handwritten Text. arXiv preprint arXiv:1907.11845, 2019. 1, 2, 3

[9] Atsunobu Kotani, Stefanie Tellex, and James Tompkin. Generating Handwriting via Decoupled Style Descriptors. In ECCV, 2020. 1, 2

[10] Eloi Alonso, Bastien Moysset, and Ronaldo Messina. Adversarial generation of handwritten text images conditioned on sequences. In ICDAR, pages 481–486. IEEE, 2019. 1, 2 , 3, 6

[11] Brian Davis, Chris Tensmeyer, Brian Price, Curtis Wiging-ton, Bryan Morse, and Rajiv Jain. Text and style conditioned gan for generation of offline handwriting lines. BMVC, 2020.1, 2, 3, 6, 7, 8, 5, 9, 10, 11, 12, 13

[12] Lei Kang, Pau Riba, Yaxing Wang, Marc¸al Rusi˜nol, Alicia Fornes, and Mauricio Villegas. Ganwriting: Content-conditioned generation of styled handwritten word images. In ECCV, pages 273–289. Springer, 2020. 1, 2, 3, 6, 7, 8, 5, 9, 10, 11, 12

[13] Sharon Fogel, Hadar Averbuch-Elor, Sarel Cohen, Shai Mazor, and Roee Litman. Scrabblegan: semi-supervised varying length handwritten text generation. In CVPR, pages 4324–4333, 2020. 1, 2, 3, 7

[14] Bo Ji and Tianyi Chen. Generative Adversarial Network for Handwritten Text. arXiv preprint arXiv:1907.11845, 2019. 1, 2

[15] Atsunobu Kotani, Stefanie Tellex, and James Tompkin. Generating Handwriting via Decoupled Style Descriptors. In ECCV, 2020. 1, 2

[16] Ji Gan and Weiqiang Wang. HiGAN: Handwriting Imitation Conditioned on Arbitrary-Length Texts and Disentangled Styles. In AAAI, 2021. 1, 2, 3, 6, 7, 8

[17] Ji Gan, Weiqiang Wang, Jiaxu Leng, and Xinbo Gao. HiGAN+: Handwriting Imitation GAN with Disentangled Representations. ACM Trans. Graphics, 42(1):1–17, 2022. 1, 2, 3

[18] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Mubarak Shah. Handwriting Transformers. In ICCV, 2021. 1, 2, 3, 4,6, 7, 8