基于对抗生成网络的多风格化的汉字

陈杰夫[1*]，陈 华[2]，徐 行[1]，姬艳丽[1]，陈李江[3]

(1. 电子科技大学计算机科学与技术学院 成都 610054；2. 江西师范大学计算机信息工程学院 南昌 330027；
3. 北京阿凡题科技有限公司 北京 海淀区 100083)

【摘要】随着生成对抗网络(GAN)的发展，中文字体转换领域的研究越来越多，研究者能够生成高质量的汉字图像。这些字体转换模型可以使用GAN将源字体转换为目标字体。然而，目前的方法有以下局限：1) 生成的图像模糊；2) 模型一次只能学习和生成一种目标字体。针对这些问题，该文开发了一种全新的模式来执行中文字体转换。首先，将字体信息附加到图像上，告诉生成器需要转换的字体；然后，通过卷积网络提取和学习特征映射，并使用转置卷积网络生成照片真实图像。使用真实图像作为监控信息，以确保生成的字符和字体与它们自身一致。这个模型只需要训练一次，就能够将一种字体转换为多种字体并生成新的字体。对7个中文字体数据集的大量实验表明，该方法在中文字体转换中优于其他几种方法。

关 键 词 中文字体样式转换；生成对抗网络；多域；创建新字体

中图分类号 TN97 **文献标志码** A doi:10.3969/j.issn.1001-0548.2019.05.003

Learning to Write Multi-Stylized Chinese Characters by Generative Adversarial Networks

CHEN Jie-fu[1*], CHEN Hua[2], XU Xing[1], JI Yan-li[1], and CHEN Li-jiang[3]

(1. College of Computer Science and Technology, University of Electronic Science and Technology of China Chengdu 610054;
2. School of Computer Information Engineering, Jiangxi Normal University Nanchang 330027;
3. Beijing Afanti Inc. Haidian Beijing 100083)

Abstract With the development of Generative Adversarial Networks (GAN), more and more researches have been conducted in the field of Chinese fonts transformation and researchers are able to generate high-quality images of Chinese characters. These font transformation models can transform a source font to a target font using GAN. However, current methods have limitations that 1) generated images are oftentimes blurry and 2) models can only learn and produce one target font at a time. To address these problems, we have developed a brand-new model to perform Chinese font transformation. First, font information is attached to images to tell the generator the fonts that we want to transform. Then, the generator extracts and learns feature mappings through convolutional networks and generates photo-realistic images using transposed convolutional networks. The ground truth images are then used as supervisory information to ensure that characters and fonts generated are consistent with themselves. This model only needs to be trained once, but it is able to transform one font to multiple fonts and produce new fonts. Extensive experiments on seven Chinese font datasets show the superiority of the proposed method over several other methods in Chinese font transformation.

Key words Chinese font styles transformation; generative adversarial networks; multiple domains; new font creation

1 Introduction

Chinese font transformation and font design have always been problematic. One problem is that, compared to English or Latin letters, the total number of Chinese characters is huge. Chinese government standard GB18030-2000, there are 27 533 unique characters[1] and the number of daily used characters is

Received date: 2019 – 07 – 01；Revised date: 2019 – 09 – 08
收稿日期：2019 – 07 – 01；修回日期：2019 – 09 – 08
Foundation item: Supported by the National Natural Science Foundation of China under Grant(61602089，61673088)
基金项目：国家自然科学基金(61602089，61673088)
Biography: CHEN Jie-fu was born in 1993, and his research interests include multimedia content analysis, computer vision and social media analysis.
作者简介：陈杰夫(1993 –)，男，主要从事多媒体内容分析、计算机视觉和社交媒体分析等方面的研究. E-mail: 790416231@qq.com

at least 3 500. Another problem is that Chinese characters have complex shapes and structures and researchers cannot simply transform them by classifying them.

Some methods, such as zi2zi and Chinese typography transfer, implement Chinese font transformation based on Pix2Pix[2], which is an image-to-image translation. These models learn feature maps from one target font and then apply the feature maps to the source font to do the transformation. However, one limitation of doing so is that we need to train this model again if we want to transform characters to another target font, which can be very time-consuming. Another problem is that the images generated are oftentimes blurry.

Another image-to-image translation research is StarGAN [3]. StarGAN can learn the mappings among multiple domains using only a single generator and a discriminator, training effectively from images of all domains.

Nevertheless, this method cannot work effectively on transforming Chinese fonts because Chinese characters have different radicals, graphic components and strokes.

To address these problems, we have developed a new GAN's method. We use only one generator and one discriminator. When given an image and a label (one-hot vector) of the font of interest to the generator, the generator will generate fonts corresponding to the given label. Then given the fake image and real image to the discriminator, the discriminator will discriminate which one is the real image and give labels to both images on the font style. If given one image and more than one labels, the generator can create a new font. We will explain the experimental content about optimization process and the loss function in detail at proposed framework.

In short, our main contributions are as follows.

1) Propose a novel Chinese font transform method and a new font creation method, the former can transform from one font to multiple fonts while the latter can combine multiple fonts to generate a new font.

2) Improve the GAN's generator, let it learn the specified multiple fonts information and specify the font to be generated.

3) Produce qualitative and quantitative Chinese characters images, compared with other methods.

2 Related Work

Chinese Fonts Transformation. Some Chinese font transformation methods[4-5] view Chinese characters as the combination of radicals and strokes. zi2zi is the first deep model. It views Chinese characters as images, uses CGAN[6] to transfer typography style[7-8], and can successfully transform fonts of Chinese characters. Style-Aware Auto-Encoder (SA-VAE) can capture different graphic components of Chinese characters by disentangling the latent features into content-related and style-related components. Chinese typography transfer is an end-to-end model which does not rely on the graphical components of Chinese characters or their stroke orders, and this model treats each single Chinese character as an inseparable image.

StarGAN. StarGAN is used for handling face image styles conversion. It can take in training data of multiple domains, and learn the mappings among all available domains using only one single generator[3]. This generator can combine learned feature mappings to generate new images.

Generative Adversarial Networks. Generative adversarial network (GAN)[9] has shown its superior performance in computer vision and image translation. A typical GAN model consists of two modules: a generator and a discriminator. The generator learns from the real samples to generate fake samples to "fool" the discriminator, and the discriminator tries to distinguish the real samples from the fake ones.

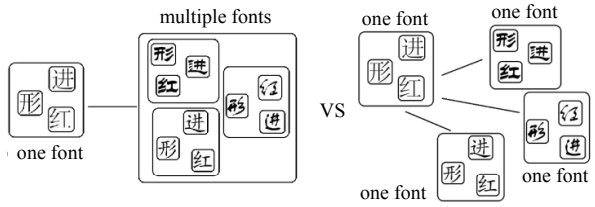
Image-to-Image Translation. In recent years with development of GAN, image-to-image translation has achieved great success in the field of image migration. For instance, Cycle-GAN[10] and DiscoGAN[11] preserve key attributes between the input and the translated images by utilizing a cycle consistency loss.

Motivated by zi2zi and StarGAN, we developed a new method to generate different fonts of Chinese

characters. Our approach is to use different font datasets and corresponding labels to train the generator. In this way, the generator can successfully generate fonts that we specify.

3 Proposed Framework

In this section, we will explain how our experimental framework works. The comparison with other methods is shown in Fig. 1.



a. Our proposed model b. Other models

Fig. 1 Our model compare with other models

Let t denote the ground truth image, x denote the input image and y denote the fake image and c denote the target font label. Let G be the generator and $G(x,c)\rightarrow y$ means combining x and c to feed into G to generate a fake image y. Suppose the discriminator is D and $D(\boldsymbol{x})\rightarrow\{D_{\text{dis}}(\boldsymbol{x}),D_{\text{cls}}(\boldsymbol{x})\}$ means giving an image $\boldsymbol{x}$ into the discriminator. D_{dis} means the result that D produces on whether the image is real. D_{cls} is the probability distribution of the target font labels that D produces.

Framework steps. In our model, we use x with one style and $\boldsymbol{c}$ as input label (one-hot vector). Then, we merge $\boldsymbol{x}$ and $\boldsymbol{c}$ into a matrix. Using the matrix feeds G and generates fake images. Finally, D discriminates the images true or false, and each image is given a font label probability.

In order to get good experimental results, we use the adversarial loss (see formula (1) and (2)) to ensure we can produce high-quality images. The semantic consistent loss (see formula (7)) is used to keep the contents of input images and output images consistent. The font classification loss (see formula (3) and (4)) helps the model to generate and transform the fonts correctly.

Adversarial loss. Below is the basic adversarial loss function to ensure that the image generated by the generator can "fool" the discriminator:

$$L_{\text{adv}}^{\text{d}}=E_{\boldsymbol{t}}[\log D_{\text{dis}}(\boldsymbol{t})]+E_{\boldsymbol{x},\boldsymbol{c}}[\log(1-D_{\text{dis}}(G(\boldsymbol{x},\boldsymbol{c})))] \quad (1)$$

Here D tries to distinguish real images from photo-realistic images which are generated by G. We use $\boldsymbol{t}$ as the supervision to make D to have distinguishing ability to the maximum extent.

This loss function is to assist G making photo-realistic images to fool the discriminator:

$$L_{\text{adv}}^{\text{g}}=E_{\boldsymbol{x},\boldsymbol{c}}[\log(1-D_{\text{dis}}(G(\boldsymbol{x},\boldsymbol{c})))] \quad (2)$$

Here G tries to minimize this objective.

Fonts Classification Loss. We give the ground truth images $\boldsymbol{t}$ and ground truth label $\boldsymbol{c}'$ as the supervision to let D learn to classify the fonts by minimizing objective below. $D_{\text{cls}}(\boldsymbol{c}'\mid\boldsymbol{t})$ is a probability distribution on target font's labels computed by D. The loss defines as

$$L_{\text{cls}}^{\text{d}}=E_{\boldsymbol{t},\boldsymbol{c}'}[-\log D_{\text{cls}}(\boldsymbol{c}'\mid\boldsymbol{t})] \quad (3)$$

The loss function for font classification of fake images is defined as

$$L_{\text{cls}}^{g}=E_{\boldsymbol{x},\boldsymbol{c}}[-\log D_{\text{cls}}(\boldsymbol{c}\mid G(\boldsymbol{x},\boldsymbol{c}))] \quad (4)$$

Here G tries to minimize this objective to generate fake images which will be classified as target labels.

Gradient penalty loss. We use gradient penalty[14] to get faster convergence and produce higher-quality photorealistic samples. $\nabla_{x'}$ denotes gradient. α is a hyperparameter. The loss is defined as

$$x'=\alpha*t+(1-\alpha)*x \quad (5)$$

$$\text{GP}_{\boldsymbol{x},\boldsymbol{t}}=\lambda_{\text{gp}}(\|\nabla_{\boldsymbol{x}'}D_{\text{dis}}(\boldsymbol{x}')\|_2-1)^2 \quad (6)$$

Semantic consistent loss. In our model, we want generated Chinese characters to be the same as the given ones, so we use the $L1$ loss function. The semantic consistent loss is defined as

$$L_{\text{feat}}=E_{\boldsymbol{x},\boldsymbol{c}}[\|\boldsymbol{t}-G(\boldsymbol{x},\boldsymbol{c})\|_1] \quad (7)$$

and minimizing this objective can make G keeping the content consistent.

Final optimization objective function. Combining all the loss functions, we train the final optimization objective function as

$$\min_G\max_D L_{DG}=L_D+L_G \quad (8)$$

and

$$L_D=\lambda_{\text{adv}}L_{\text{adv}}^{\text{d}}+\lambda_{\text{cls}}(L_{\text{cls}}^{\text{d}}+GP_{\boldsymbol{x},\boldsymbol{t}}) \quad (9)$$

$$L_G=\lambda_{\text{adv}}L_{\text{adv}}^{\text{g}}+\lambda_{\text{cls}}L_{\text{cls}}^{\text{g}}+\lambda_{\text{feat}}L_{\text{feat}} \quad (10)$$

Where λ_{adv}, λ_{feat} and λ_{cls} are weights that apply to losses to have a better trade-off in semantics, classification and adaptation.

Algorithm. The algorithm of the proposed method is as follows.

Input: Source image $\boldsymbol{x}$ and target label $\boldsymbol{c}$; Target image $\boldsymbol{t}$ and target label $\boldsymbol{c}$

randomly initialized a generator G and a discriminator D

repeat

for number of training epochs do

for number of batch-size do

//for generator

$$\theta_G \leftarrow \theta_G - \mu \frac{\partial L_{DG}}{\partial \theta_G},\ L_{DG} \text{ as Eq.7}$$

//for discriminator

$$\theta_{D_{\text{dis}}} \leftarrow \theta_{D_{\text{dis}}} + \mu \frac{\partial L^{\text{d}}_{\text{adv}}}{\partial \theta_{D_{\text{dis}}}}, L^{\text{d}}_{\text{adv}} \text{ as Eq.1}$$

$$\theta_{D_{\text{cls}}} \leftarrow \theta_{D_{\text{cls}}} + \mu \frac{\partial L^{\text{d}}_{\text{cls}}}{\partial \theta_{D_{\text{dis}}}}, L^{\text{d}}_{\text{cls}} \text{ as Eq.3}$$

$$\theta_D \leftarrow \theta_{D_{\text{dis}}} + \theta_{D_{\text{cls}}}$$

end for

end for

Until convergence

$$\hat{\theta}_D \leftarrow \theta_D$$

$$\hat{\theta}_G \leftarrow \theta_G$$

Output: the optimized G and D by $\hat{\theta}_D, \hat{\theta}_G$

4 Experimental and Results

4.1 Experimental Settings

For comparing with zi2zi and Chinese typography transfer, the font2img script from zi2zi was used to generate Chinese font datasets. The seven Chinese font datasets include: Songti style, Heiti style, Kaiti style, Lishu style, Xinwei style, Shuti style, and Xingkai style. 3 498 Chinese characters were generated for each dataset, generating total of 24 486 Chinese character images. The image size of each character is set to 64×64×3. A number to each font is assigned as the label, i.e., 0 to Songti style, 1 to Heiti style, 2 to Kaiti style, 3 to Lishu style, 4 to Xinwei style, 5 to Shuti style, and 6 to Xingkai style. After that we converted each label to one-hot label.

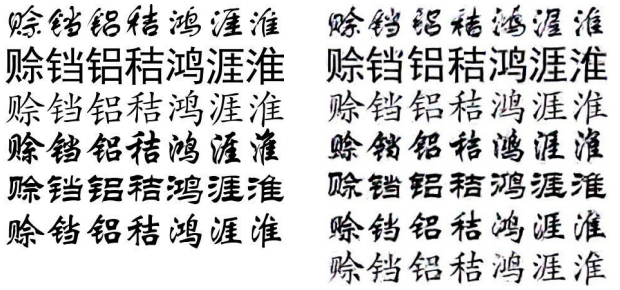
To let the generator know the font we specified, input images and labels is attached together. As a result, the input size of images is changed to 64×64×(3+7) while the output size remains as 64×64×3.

4.2 Network Architecture

Inspired from StarGAN and ADDA[12], our model has one generator and one discriminator, which has the capabilities of discrimination and classification. The generator network is composed of three convolutional blocks, six residual blocks[13], and three transposed convolutional blocks. The instance normalization is used for each convolutional layer and a rectified linear unit (Relu) activation function is followed. The discriminator network is composed of six convolutional blocks and two fully convolutional networks. Each convolutional block contains a convolutional layer and a Leaky-Relu activation function. One of the fully convolutional networks is used to distinguish real and fake images, and the other ones are used to classify the fonts.

4.3 Results on zi2zi's Dataset

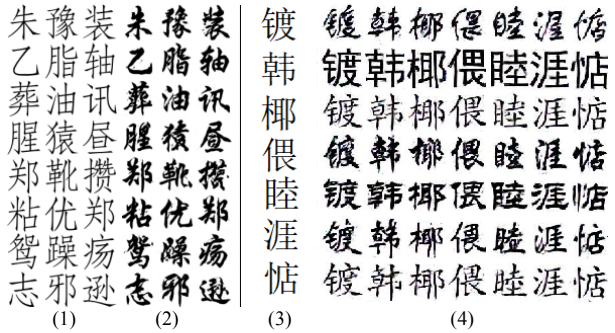
Choosing one font as the source font and the rest six fonts as the targets, we produced high-quality images with fonts specified. The real images and photo-realistic images generated by using Songti Style are shown in Fig. 2. From the figure we can see that the produced images have clear graphical structures and strokes. It can be seen the photo-realistic image in last line generated by combing Xingkai and Xinwei Style is very successful.



a. Real images　　b. Photo-realistic images

Fig. 2　Real images and our generating images

Comparing with other models which can only produce one font, our model can successfully not only generate six target fonts by using one source font, but also create one new font by combining any fonts, and all fonts generated have clear images. The comparing results are shown in Fig. 3 and Fig. 4.



(1) (2) (3) (4)

a. Chinese typography transfer method, one to one transform, i.e. font (1) to font (2)

b. Our method, one to multiple transform, i.e. font (3) to fonts (4) which has six different fonts

Fig. 3 Our model′s results compare with other models′ results

In theory, the ability of combining fonts makes it possible for our model to generate 63 new fonts.

a. Input image (Songti Style)

b. zi2zi results (Only one font can be generated at a time)

c. Our results (Multiple fonts can be generated at a time

(our result)

(new)

Fig. 4 Comparison of the results of three models

5 Conclusion

In this paper, we proposed a new model to do one-to-many Chinese fonts transformation and to produce new fonts by combining existing fonts. Comparing with zi2zi and Chinese typography transfer, our model can produce higher-quality images and is reusable for different fonts. This reusable feature saves a lot of time compared to modifying and training the model again. Besides, the reuse of the same model is a major focus of transfer learning for deep learning.

References

[1] SUN Dan-yang, REN Tong-zheng, LI Chong-xuan, et al. Learning to write stylized Chinese characters by reading a handful of examples[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm, Sweden: [s.n]. 2018: 920-927.

[2] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C]//IEEE Conference on Computer Vision & Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 5967-5976

[3] CHO Yun-jey I, CHOI Min-je, KIM Mun-young, et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation[C]//IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 8789-8797.

[4] XU Song-hua, JIN Tao, JIANG Hao, et al. Automatic generation of personal Chinese handwriting by capturing the characteristics of personal handwriting[C]//Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence. Pasadena, California, USA: [s.n.], 2009.

[5] XU Song-hua, JIANG Hao, JIN Tao, et al. Automatic generation of Chinese calligraphic writings with style imitation[J]. IEEE Intelligent Systems, 2009, 4(2): 44-53.

[6] MIRZA M, OSINDERO S. Conditional generative adversarial nets[EB/OL]. [2014-11-06]. https://arxiv.org/abs/1411.1784.

[7] SUN Han-fei, LUO Yi-ming, LU Zhang. Unsupervised Typography Transfer[EB/OL]. [2018-02-07]. https://arxiv.org/abs/1 802.02595.

[8] CHANG Jie, GU Yu-jun, ZHANG Ya. Chinese typography transfer [EB/OL]. [2017-07-16]. https://arxiv.org/abs/1707.04904.

[9] GOODFELLOW I J, JEAN P A, MIRZA M, et al. Generative adversarial networks[EB/OL]. [2014-06-04]. https://arxiv.org/abs/1 406.2661v1.

[10] ZHU Jun-yan, PARK T, ISOLA P. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 2242-2251.

[11] KIM T, CHA M, KIM H, et al. Learning to discover cross-domain relations with generative adversarial networks[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: [s.n.], 2017: 1857-1865.

[12] TZENG E, HOFFMANM J, SAENKO K, et al. Adversarial discriminative domain adaptation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 2962-2971.

[13] HE Kai-ming , ZHANG Xiang-yu , REN Shao-qing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 770-778.

编 辑 蒋 晓