Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 2717322, 8 pages
https://doi.org/10.1155/2022/2717322

Research Article

Variational Autoencoder for Zero-Shot Recognition of Bai Characters

Weiwei Lin,[1,2] **Tai Ma**,[3] **Zeqing Zhang**,[4,5] **Xiaofan Li**,[5] **and Xingsi Xue**[6]

[1]*School of Big Data and Artificial Intelligence, Fujian Polytechnic Normal University, Fuqing 350300, China*
[2]*Engineering Research Center for ICH Digitalization and Multi-Source Information Fusion, Fujian Province University, Fuqing 350300, China*
[3]*Department of Physics, West Yunnan University of Applied Sciences, Dali 671000, China*
[4]*Department of Earth Science and Engineering, West Yunnan University of Applied Sciences, Dali 671000, China*
[5]*Department of Information, Xiamen University, Xiamen 361005, China*
[6]*Fujian Provincial Key Laboratory of Big Data Mining and Applications, Fujian University of Technology, Fuzhou, Fujian 350118, China*

Correspondence should be addressed to Zeqing Zhang; 313460472@qq.com

Received 16 May 2022; Accepted 20 June 2022; Published 4 July 2022

Academic Editor: Fuquan Zhang

When talking about Bai nationality, people are impressed by its long history and the language it has created. However, since fewer people of the young generation learn the traditional language, the glorious Bai culture becomes less known, making understanding Bai characters difficult. Based on the highly precise character recognition model for Bai characters, the paper is aimed at helping people read books written in Bai characters so as to popularize the culture. To begin with, a data set is built with the support of Bai culture fans and experts. However, the data set is not large enough as knowledge in this respect is limited. This makes the deep learning model less accurate since it lacks sufficient data. The popular zero-shot learning (ZSL) is adopted to overcome the insufficiency of data sets. We use Chinese characters as the seen class, Bai characters as the unseen class, and the number of strokes as the attribute to construct the ZSL format data set. However, the existing ZSL methods ignore the character structure information, so a generation method based on variational autoencoder (VAE) is put forward, which can automatically capture the character structure information. Experimental results show that the method facilitates the recognition of Bai characters and makes it more precise.

1. Introduction

With a population of over 1 million, Bai nationality boasts a splendid culture. Its history can date back to ancient times. The Bai people mainly settle in Dali Bai Autonomous Prefecture, Yunnan province. Besides, they also reside in Bijie Prefecture of Guizhou, Liangshan Prefecture of Sichuan, etc. Bai people communicate in their own language which serves as an emotional bond and a major carrier of Bai culture. This is the most basic national characteristic of Bai nationality. The academic circles both at home and abroad have been focusing on the linguistic structure of Bai characters whose vocabulary, pronunciation, and grammar belong to the Tibeto-Burmese. Since the literature concerning the Bai nationality is limited, the historical and cultural significance of the study is obvious.

The objective of the study is to revive Bai culture and decipher the Bai characters, helping people to understand historical literature of Bai people or inscriptions of Bai characters. Therefore, a model recognizing Bai characters should be established so that when an unknown Bai character appears, it can be recognized and explained by the model.

As the neural network becomes popular, great achievements have been made on visual tasks [1, 2]. Therefore, deep learning models are proposed [1, 3–5] to recognize Bai characters one by one. Unfortunately, the data hungry nature of convolutional networks leads to a significant decline in their performance when there is less training data. At the same

time, the collection of Bai character data set needs a lot of expert knowledge and is expensive, so the established data set is not as large as the traditional data set for classification [6, 7]. To sum up, a single data set of Bai characters alone makes the training of a deep learning model difficult and the ideal results are hard to get.

To this end, we consider using zero-shot learning (ZSL) [8, 9] as collecting Bai character data sets is hard to obtain [10]. ZSL can transfer knowledge from visible categories to invisible ones according to their attributes to prevent samples of the invisible category from decreasing or even losing. If we treat Bai characters as unseen classes, we need to collect data sets similar to Bai characters as seen classes in order to better transfer knowledge [10]. Fortunately, it can be seen that Chinese and Bai characters are highly similar as they come from the Sino-Tibetan [11, 12] language family, as shown in Figure 1. In addition, Chinese character data sets are very easy to collect and cheap and do not need expert knowledge. Therefore, we collected a huge Chinese character data set as seen class in ZSL. Last but not least, ZSL needs to depict the attributes of both visible and invisible categories, so as to better transfer knowledge through attributes [10]. It is found that characters of Chinese and Bai are made up of 32 basic strokes, as shown in Figure 2. The case is similar in English which is made up of 26 letters. As a result, it is reasonable to regard the number of different strokes of each word as an attribute.

After building the data set in ZSL format, we transplanted some classical methods in ZSL, which include projection methods: DAP [9] and IAP [9], and generation method- [13] based GAN [14]. These methods establish the relationship between attributes and features, so as to generalize Chinese characters to Bai characters and finally improve the accuracy [10]. However, our attributes only contain the stroke information of characters, so these models ignore the important information of character structure in the process of training.

As a possible solution, a generation method built on VAE [15] is introduced in this paper. VAE includes an encoder and a decoder. The former can obtain the semantic information of the characters, and the latter can reconstruct the characters according to the information and attributes. The reconstruction of characters obviously needs to consider the structure of characters. Therefore, thanks to the reconstruction loss, VAE can automatically capture the structure information of the characters, making the character features synthesized by the decoder more realistic. Experiments show that the generation method based on VAE has achieved amazing accuracy improvement, which is far better than the classical ZSL methods.

The significance of the study is fourfold: (1) a data set in ZSL format with Bai language and Chinese is built. (2) A generation method built on VAE is proposed, which can automatically capture the structure information of characters. (3) According to experimental results, the method enhances the correction rate of recognizing Bai characters.

2. Related Work

2.1. Text Recognition of Bai Characters. Zhang et al. [10] as a pioneer resorting to deep learning to recognize the Bai text collected a data set with 400 words of Bai handwritten fonts. In their work, Zhang et al. [10] considered that the Chinese language is similar to Bai language to facilitate the model to recognize Bai characters based on learning transfer [10, 16] and achieved remarkable results. However, their study is limited to a limited data set, and the data set contains only 400 Bai characters. Once there are Bai characters that are not in the data set, their trained model is bound to obtain a wrong classification result. In contrast, we consider using GZSL in such a way that the model gains the ability to recognize Bai characters outside the training set, which greatly improves the application value of the model.

2.2. Zero-Shot Learning. Currently, generative approaches dominate in GZSL, which exploit existing adversarial generative networks (GAN) [14, 17, 18] or variational autoencoders (VAE) [15, 19, 20] so that visual characteristics from class-level semantic attributes and random noise can be synthesized. f-CLSWGAN [13], cycle-UWGAN [21], and LisGAN [22] introduce the Wasserstein generative adversarial network (WGAN) [23] coupled with a pretrained classifier so that visual characters for invisible characteristics can be synthesized, thus allowing the GZSL work to deteriorate into a fully supervised issue for categorization. RFF [24] combines the traditional projection method and GAN to initially map visual characteristics to a new feature space without redundancy and judge the veracity of the mapped characteristics. Different from methods based on GAN, some works [25–28] formulate GAN in the variational autoencoder (VAE) model to match the latent distribution that is based on categories and feature representations that are discriminating. They further improve the quality of synthetic features by combining the advantages of GAN and VAE and even extend the model to the transductive setting through an unconditional discriminator. However, all these methods only optimize their own models on seen classes and consider how a good generator can be trained to synthesize visual characteristics according to attributes, without directly simulating zero-shot learning settings so that knowledge can be transferred from visible categories to invisible ones during training.

To better mimic the zero-shot learning settings, previous studies [29–31] introduce meta-learning to make the model more suitable for transferring knowledge from seen classes to unseen classes. ZSML [30] combines metalearning with GAN and utilizes a single gradient update to obtain a generic initialization suitable for internal learning. E-PGN [29] first introduces an episode-based paradigm for training, which trains the model by simulating multiple zero-shot classification tasks on the seen classes. After training a collection of episodes, the model is expected to be an expert in predicting unseen classes such that it can generalize well to the real unseen classes. TGMZ [31] proposed a task-wise model, which extracts both class and visual feature information for reconstruction, to carry out task-wise distribution alignment.

3. Method

3.1. Problem Definition. The zero-shot learning (ZSL) [9] methods are introduced to recognize Bai characters. In the recognition task, the data set is made up of two disjoint sets:

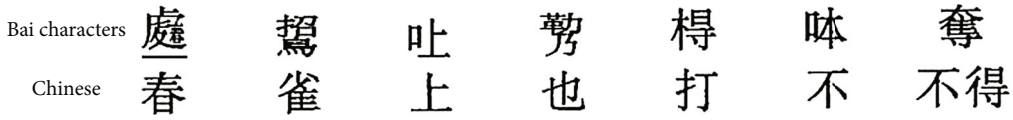


Figure 1: Contrast between Bai characters and Chinese characters.

the Chinese character set and the Bai character one, which are regarded as the set of visible and invisible categories in the ZSL working, respectively. Suppose that $D_s = \{X_s, Y_s, A_s\}$ denotes the Chinese character set, and $D_u = \{X_u, Y_u, A_u\}$ denotes the Bai character set. X_s and X_u are the instance set, which are for training and testing. Y_s and Y_u belong to the label sets. A_s and A_u are the corresponding attribute sets constructed from strokes. Note that $Y_s \cap Y_u = \varnothing$, which means that the categories involved in training and testing are disjoint. In the study, the Chinese character set and the attributes of Bai character set are adopted to train the classifier which can recognize the Bai characters. This is of great significance for the recognition of Bai characters, which are difficult to collect. The current ZSL methods will be introduced below, and they will be applied to Bai character recognition.

3.2. Intermediate Attribute Classifier for Learning. Direct Attribute Prediction (DAP) [9] and Indirect Attribute Prediction (IAP) [9] are the earliest ZSL methods which make use of the attributes to infer the label of the instances. We introduced DAP and IAP to the Bai character recognition task. For simplicity, suppose that the attribute representations $a^c = (a_1^c, \cdots, a_c^M)$ of classes for training (Chinese character classes) and c are fixed-length vectors.

The probabilistic classifier is learned for the attribute a_m based on DAP. As shown in Figure 3(a), the trained classifier is estimated to be $p(a_m|x)$. Then, a model showing the entire image-attribute layer as $p(a|x) = \prod_{m=1}^{M} p(a_m|x)$ is established. During the test, every invisible category (Bai character classes) induces its attribute vector a^b deterministically, i.e., $p(a|b) = \mathrm{I}(a = a^b)$, where $I(P)$ is the indicator function: $I(P) = 1$ if P is true; then, it is 0 otherwise. Based on the principles of Bayes, the attribute-class layer is as follows: $p(b|a) = (p(b)/p(a^b))I(a = a^b)$. Combining image-attribute and attribute-class layers, when X is given, the posterior can be obtained by

$$p(b|x) = \sum_{m=1}^{M} p(b|a_m)p(a_m|x) = \frac{p(b)}{p(a^c)} \prod_{m=1}^{M} p\left(a_m^b|x\right). \quad (1)$$

When there is not enough specific knowledge, the factor $p(b)$ below can be ignored. Regarding the factor $p(a)$, $p(a) = \prod_{m=1}^{M} p(a_m)$ is assumed. The empirical means $p(a_m) = (1/K) \sum_{k=1}^{K} a_m^c$ is adopted for classes for training as prior attribute. Based on the principle of decision $f : X \longrightarrow B$ optimizing output class based on test classes $b_1, \cdots, b_L$ to a test sample x, it

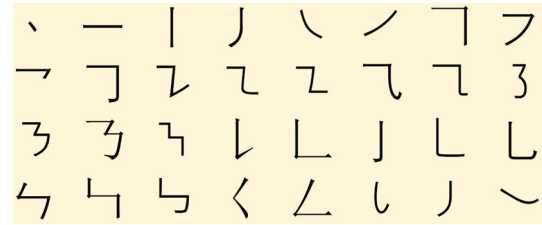


Figure 2: 32 characters of Chinese and Bai, basic strokes.

can be predicted that

$$f(x) = \arg \max_b \prod_{m=1}^{M} \frac{p\left(a_m^b|x\right)}{p\left(a_m^b\right)}. \quad (2)$$

According to IAP, the probabilistic attributes of input x are indirectly estimated by first predicting the probabilities of every class; then, the attribute matrix is multiplied as shown in Figure 3(b). The attribute probabilities are computed by

$$p(a_m|x) = \sum_{k=1}^{K} p(a_m|c_k)p(c_k|x). \quad (3)$$

where $p(a_m|c_k)$ is the attribute of the class that is defined beforehand and $p(c_k|x)$ is the (Chinese character) posterior of the class. After $p(a_m|x)$ has been computed, the class label of testing is predicted based on Equation (2) (Bai character).

3.3. Feature Generative Framework

3.3.1. Generative Adversarial Network. The generative adversarial network (GAN) [13, 14] proposes the feature generative framework for zero-shot learning. We introduce it into zero-shot Bai character recognition. Feature generative framework is divided into two stages: feature generation and classification. As shown in Figure 4(a), for the GAN-based feature generative framework, in feature generation, a conditional generator G is trained to synthesize the samples $\tilde{x} = G(a, z)$ when a Gaussian noise $z \in N(0, 1)$ and the attribute a are taken into consideration. The discriminator D is crossiteratively trained with the generator and learns to distinguish a real pair (x, a) from a synthetic one $(\tilde{x}, a)$. G attempts to synthesize a more realistic feature $\tilde{x}$ to confuse the discriminator during training. Besides, the generator hopes to find the right synthetic feature $\tilde{x}$ to its corresponding attribute a. The generative model adopts the structure of WGAN [32] and introduces the gradient penalty term [23];

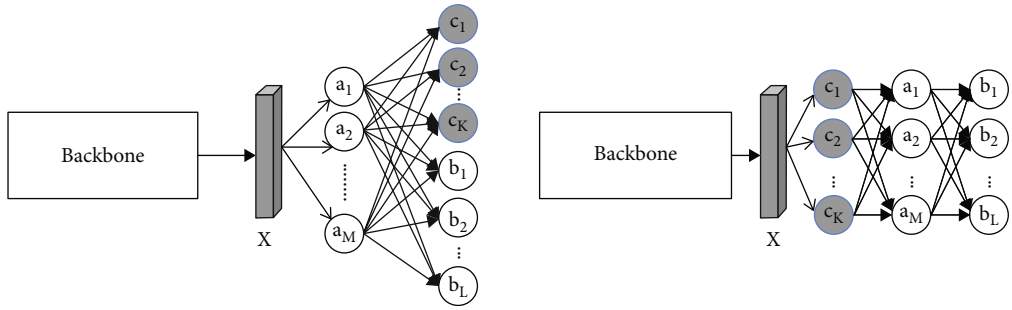


Figure 3: The model structure diagram for DAP and IAP. x is the feature vector extracted by the backbone. $(a_1, a_2, \cdots, a_M)$ is the attribute vector. $(c_1, c_2, \cdots, c_K)$ and $(b_1, b_2, \cdots, b_L)$ represent the label of Chinese and Bai character, respectively. The thin arrow shows trainable weights, and thick arrow lines show predefined weights.

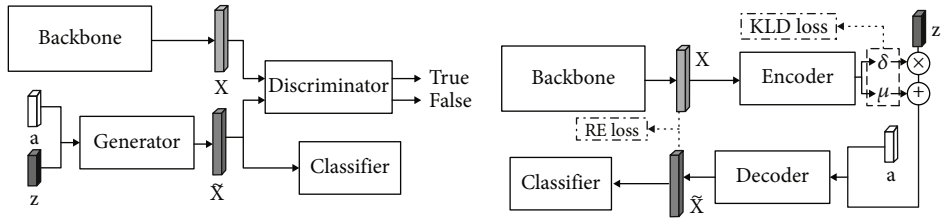


Figure 4: The feature generative framework based on GAN and VAE. x is the visual feature extracted by the backbone. $\tilde{x}$ is the synthetic feature. a is the feature of character, and z is a Gaussion noise. δ and μ are the mean and variance required for reparameter, respectively.

Figure 5: Some examples of Bai character data set.

Table 1: Methods compared based on accuracy. Bai characters are not used as a direct way to train the model.

Method	AlexNet	VGG19	ResNet101
DAP [33]	46.28	49.92	53.84
IAP [33]	42.23	47.26	48.59
GAN [13]	67.54	70.78	76.46
VAE (ours)	70.47	74.81	78.14

the adversarial training loss of G and D is as follows:

$$\mathscr{L}_{\text{wgan}} = \mathrm{E}[D(x,a)] - \mathrm{E}[D(\tilde{x},a)] - \lambda \mathrm{E}\left[\left(\|\nabla \hat{x}, a\|_2 - 1\right)^2\right]. \quad (4)$$

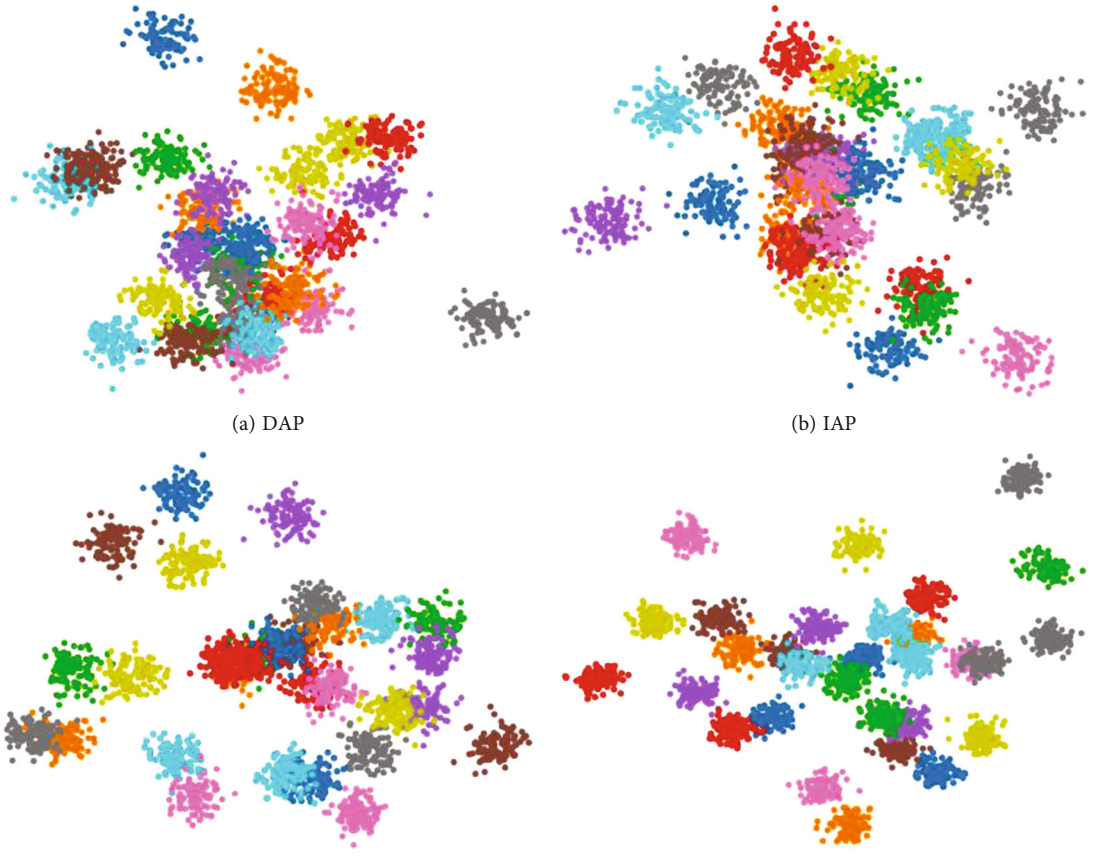
The generator is trained on the Chinese character set, and the trained generator is used to synthesize Bai character samples. In the stage of classification, synthetic samples are used to train the classifier to recognize Bai character through the crossentropy loss:

$$\mathscr{L}_{\text{ce}} = \mathrm{E}[\log P(b|x;\theta)], \quad (5)$$

where $\theta \in \mathrm{R}^{d_x \times L}$ is the weight matrix of an interconnected layer. d_x is the dimension of x, and L is the class number of Bai character, and $P(b|x;\theta) = \exp(\theta_b^T x)/\sum_i^L \exp(\theta_i^T x)$. The prediction function is

$$f(x) = \arg \max_b P(b|x;\theta). \quad (6)$$

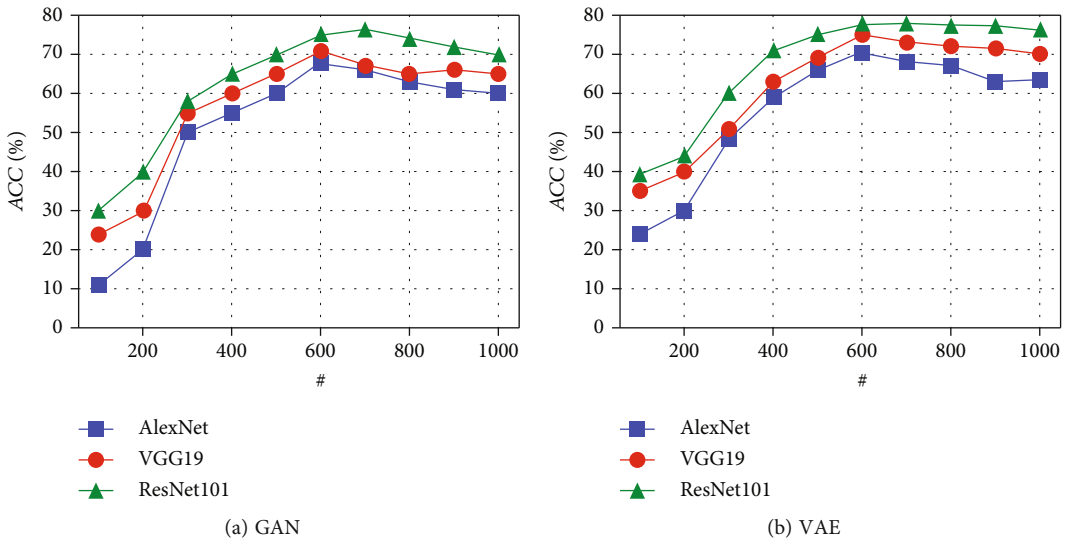
3.3.2. Variational Autoencoder. In addition to the GAN-based generative framework, the variational autoencoder (VAE) [15] sets the base for the proposed generative framework, which is more suitable for character recognition task. According to Figure 4(b), without discriminator, the framework based on VAE is made up of the encoder E_n and decoder D_e (generator). During the training, the former uses reparameter trick and is trained through the Kullback-Leibler Divergence (KLD) loss which is demonstrated as



(a) DAP (b) IAP

(c) GAN (d) VAE

Figure 6: Visualization results of Bai character features by different methods.



(a) GAN (b) VAE

Figure 7: The performance of GAN-based and VAE-based generative methods with numbers of synthetic characteristic for each Bai character class. ACC is the average top-1 accuracy for each class. # is the number of synthetic features per Bai character class.

follows:

$$\mathscr{L}_{\text{kld}} = \mathrm{E}\left[\frac{1}{2}\sum_{i=1}^{N_d}\left(1 + \log\left(\delta_i^2 - \mu_i^2 - \delta_i^2\right)\right)\right], \tag{7}$$

where μ and δ are the outputs of the encoder, representing the mean and variance, respectively. N_d refers to the dimension of μ and δ. Through reparameter, hidden variables encoded by the encoder can be represented as

$$\tilde{z} = z \times \delta + \mu. \tag{8}$$

At the same time, the decoder D_e is trained together with the encoder E_n through the reconstruction (RE) loss:

$$\mathscr{L}_{\text{re}} = \mathrm{E}\left[\|x - \tilde{x}\|_2^2\right], \tag{9}$$

where $\tilde{x} = D_e(a, \tilde{z})$ and the RE loss can make the synthetic feature $\tilde{x}$ have better structural information. Thereby, compared with GAN, VAE is more suitable for character recognition which pays more attention to structure. In feature generation, the total loss of VAE is as follows:

$$\mathscr{L}_{\text{vae}} = \mathscr{L}_{\text{re}} + \lambda\mathscr{L}_{\text{kld}}, \tag{10}$$

where λ refers to a hyperparameter. After training, the decoder is used to synthesize visual features for Bai character conditioned the attributes and Gaussian noises. To synthesize features of Bai character which are leverage to train the classifier through Equation (5) and Equation (6) is used for prediction.

4. Experiment

4.1. Experimental Setup

4.1.1. Data Sets. A big data set of Bai characters is built (see Figure 5). The data set includes 400 Bai characters. Since some characters in Bai and Chinese overlap, those Bai characters differing from Chinese ones are included in the data set. Each word has 50 samples, and they are written by Bai people and Bai culture fans. Besides, a data set of sufficient Chinese characters is also chosen, including 509 Chinese characters. There are about 1,000 samples for each character. To build a data set in ZSL format, the two data sets are labelled with class-level attributes (a number of 32 basic strokes). The Chinese one is used to train, and the Bai character one is used to test.

4.1.2. Evaluation Protocol. The proposed method is assessed based on the average top-1 accuracy (ACC) for each class.

4.1.3. Classification Model. Three backbones, AlexNet [4], VGG19 [5], and ResNet101 [1], are adopted. Multiple models are compared; it can be analyzed that the method is suitable for all models and can be generalized.

4.2. Accuracy Analysis. The accuracy comparison of training strategies is shown in Table 1. The accuracy of DAP, IAP, GAN, and our proposed VAE on zero-shot recognition of Bai characters is assessed. Meanwhile, we evaluated the performance of the above four methods with AlexNet, VGG, and ResNet101 as the backbone, respectively. The overall experimental results show that the zero-shot learning methods will be effectively transferred for recognizing Bai characters.

DAP and IAP are projection methods. Because only Chinese character data sets are used in training, the model has never seen Bai characters. Therefore, although the model has been improved due to knowledge transfer, it is still not ideal. However, it also shows that using ZSL for knowledge transfer can transfer the knowledge of large-scale Chinese character data sets to Bai characters, so that the model can obtain a high recognition rate for Bai characters even without Bai character data as training.

GAN is a generation method. Because the generator synthesizes a large number of Bai character features, it greatly alleviates the problem of missing Bai character data and finally greatly improves the accuracy. This shows that although Chinese characters are different from Bai characters, through the attribute of stroke, the generator can well restore the features of Bai characters when only Chinese characters are trained.

VAE is our proposed method and is also a generation method. It not only has the ability to generate features but also has the ability to reconstruct the original picture. This ability plays a very important role in the transfer of text knowledge, which makes our method have the highest accuracy.

4.3. Visualization. To further illustrate the effect of these methods, the visual features obtained by different methods are visualized. After that, t-SNE [34] algorithm is adopted to finish the task (see Figure 6). DAP and IAP denote the visualization results of Bai character visual features extracted from backbone trained by DAP and IAP, respectively. GAN and VAE denote the visualization results of Bai character visual features synthesized by GAN and VAE. With the support of huge Chinese character data set, the Bai character features extracted by the network trained by DAP and IAP have been highly distinguishable, which fully shows the effectiveness of knowledge transfer. The Bai character features synthesized by the generator trained by the GAN further increased the discriminability. Finally, the Bai language features synthesized by the network trained by VAE have been well distinguishable, which fully shows the superiority of VAE for zero-shot recognition of Bai characters.

4.4. Hyperparameter Analysis. In Figure 7, we report the results of two generative methods with different numbers of synthetic features per Bai character class. It can be observed that generative methods require certain numbers of synthetic samples to achieve the desired results. But that does not mean that more synthetic samples are better. Synthetic samples create too much noise, which limits the performance of classification. Thereby, in order to obtain the ideal results of Bai character recognition, we need to limit the number of synthetic samples per Bai character class to a certain interval.

5. Conclusions

Bai nationality can date back to ancient China as a nation. It boasts of its own language and wonderful culture. However, fewer young people can read in Bai language with the passing of time and its once splendid culture is on the verge of extinction. To help Bai culture lovers and experts to read Bai literature without difficulty, the study focuses on the training of a high-precision Bai model to recognize Bai characters. Firstly, a data set of Bai characters is established. However, its size is not big enough since expert knowledge is limited. Therefore, deep learning models requiring a large amount of data cannot produce perfect results based on this data set. As a solution, the zero-shot learning (ZSL) is suggested to overcome the lack of data sets. We use Chinese characters as the seen class, Bai characters as the unseen class, and the number of strokes as the attribute to construct the ZSL format data set. However, the existing ZSL methods ignore the characters structure information, so a VAE-based generation method is proposed, which can automatically capture the character structure information. According to experimental results, the proposed methods can enhance the model to recognize Bai characters more accurately.

Data Availability

We build a large data set of Bai characters; there are a total of 400 Bai characters.

Conflicts of Interest

It is declared that no conflicts of interest exist in the study.

Acknowledgments

This work is supported by the Natural Science Foundation of Fujian Province, China (Nos. 2019J01889 and 2020J018751); the "Tiancheng Huizhi" Innovation and Education Promotion Fund, China (No. 2018A02005); and the National Natural Science Foundation of China (No. 62172095).

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, https://arxiv.org/abs/1512.03385.

[2] Y. Yan, J. Ren, G. Sun et al., "Unsupervised image saliency detection with Gestalt-laws guided optimization and visual attention based refinement," *Pattern Recognition*, vol. 79, no. 2018, pp. 65–78, 2018.

[3] Z. Fang, J. Ren, S. Marshall, H. Zhao, S. Wang, and X. Li, "Topological optimization of the DenseNet with pretrained-weights inheritance and genetic channel selection," *Pattern Recognition*, vol. 109, no. 2021, article 107608, 2021.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1106–1114, 2012.

[5] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very deep convolutional networks for natural language processing," 2016, https://arxiv.org/abs/1606.01781.

[6] Y. LeCun, C. Cortes, and C. J. C. Burges, "The MNIST database of handwritten digits," 2010, http://yann.lecun.com/exdb/mnist/.

[7] M. Swofford, "Image completion on CIFAR-10," 2018, https://arxiv.org/abs/1810.03213.

[8] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly," 2017, https://arxiv.org/abs/1707.00600.

[9] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 951–958, Miami, FL, USA, 2009.

[10] Z. Zhang, C. Lee, Z. Gao, and X. Li, "Basic research on ancient Bai character recognition based on mobile APP," *Security and Communication Networks*, vol. 2021, Article ID 4059784, 7 pages, 2021.

[11] T. Dolkar, *Sino-Tibetan relations 1990-2000: the Internationalisation of the Tibetan issue, [P.h.D. thesis]*, NA Marburg, 2008.

[12] Q. Tian and D. Jiang, "Sino-Tibetan language data and the origin of East-Asian people," *Data Science Journal*, vol. 6, pp. S715–S722, 2007.

[13] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5542–5551, Salt Lake City, Utah, 2018.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[15] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013, https://arxiv.org/abs/1312.6114.

[16] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferableare features in deep neural networks?," *Advances in Neural Information Processing Systems*, vol. 27, pp. 3320–3328, 2014.

[17] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, https://arxiv.org/abs/1411.1784.

[18] T. D. Nguyen, T. Le, H. Vu, and D. Q. Phung, "Dual discriminator generative adversarial nets," 2017, https://arxiv.org/abs/1709.03831v1.

[19] C. Doersch, "Tutorial on variational autoencoders," 2016, https://arxiv.org/abs/1606.05908.

[20] I. Higgins, L. Matthey, A. Pal et al., *Beta-VAE: learning Basic Visual Concepts with a Constrained Variational Framework*, ICLR (Poster), 2017.

[21] R. Felix, B. G. V. Kumar, I. D. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," 2018, https://arxiv.org/abs/1808.00136.

[22] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, and Z. Huang, "Leveraging the invariant side of generative zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7402–7411, Long Beach Convention Center in Long Beach, CA, 2019.

[23] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, pp. 214–223, International Convention Centre, Sydney, Australia, 2017.

[24] Z. Han, Z. Fu, and J. Yang, "Learning the redundancy-free features for generalized zero-shot object recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12862–12871, 2020.

[25] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "F-VAEGAN-D2: a feature generating framework for any-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and*

Pattern Recognition (CVPR), pp. 10275–10284, Long Beach Convention Center in Long Beach, CA, 2019.

[26] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero-and few-shot learning via aligned variational autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8247–8255, Long Beach Convention Center in Long Beach, CA, 2019.

[27] P. Ma and X. Hu, "A variational autoencoder with deep embedding model for generalized zero-shot learning," *AAAI*, vol. 34, no. 7, pp. 11733–11740, 2020.

[28] R. Keshari, R. Singh, and M. Vatsa, "Generalized zero-shot learning via over-complete distribution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13300–13308, 2020.

[29] Y. Yu, Z. Ji, J. Han, and Z. Zhang, "Episode-based prototype generating network for zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14032–14041, 2020.

[30] V. K. Verma, D. Brahma, and P. Rai, "A meta-learning framework for generalized zero-shot learning," 2020, https://arxiv.org/abs/1909.04344.

[31] Z. Liu, Y. Li, L. Yao, X. Wang, and G. Long, "Task aligned generative meta-learning for zero-shot learning," *AAAI*, vol. 35, pp. 8723–8731, 2021.

[32] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," 2017, https://arxiv.org/abs/1704.00028.

[33] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2014.

[34] N. Rogovschi, J. Kitazono, N. Grozavu, T. Omori, and S. Ozawa, "t-distributed stochastic neighbor embedding spectral clustering," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1628–1632, Anchorage, AK, USA, 2017.