Mathematical Biosciences and Engineering

MBE, 20(8): 14081–14095.
DOI: 10.3934/mbe.2023629
Received: 25 April 2023
Revised: 29 May 2023
Accepted: 04 June 2023
Published: 25 June 2023

http://www.aimspress.com/journal/MBE

Research article

Zero-shot learning via visual-semantic aligned autoencoder

Tianshu Wei[1], **Jinjie Huang**[1,2,*] **and Cong Jin**[1]

[1] School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150006, China

[2] School of Automation, Harbin University of Science and Technology, Harbin 150006, China

* **Correspondence:** Email: jjhuangps@163.com.

Abstract: Zero-shot learning recognizes the unseen samples via the model learned from the seen class samples and semantic features. Due to the lack of information of unseen class samples in the training set, some researchers have proposed the method of generating unseen class samples by using generative models. However, the generated model is trained with the training set samples first, and then the unseen class samples are generated, which results in the features of the unseen class samples tending to be biased toward the seen class and may produce large deviations from the real unseen class samples. To tackle this problem, we use the autoencoder method to generate the unseen class samples and combine the semantic features of the unseen classes with the proposed new sample features to construct the loss function. The proposed method is validated on three datasets and showed good results.

Keywords: generalized zero-shot learning; conventional zero-shot learning; autoencoder; generated samples; modalities alignment

1. Introduction

Deep learning has significantly succeeded in image recognition, image segmentation and target detection. However, deep learning models often require a large amount of labeled data to train, and labeling the data will take more time. Some scholars propose zero-shot learning. Zero-shot learning uses the model obtained from the seen classes to infer the unseen class of samples, which can largely reduce the labeling of samples [1,2]. Zero-shot learning mimics the process of human cognition of new things. For example, a person knows an animal, like a horse, through pictures and their linguistic descriptions; when knowing about the linguistic description of zebras, zebras look like horses and have black and white stripes on its body, then the person can still recognize zebras by this linguistic description when seeing zebras even though he or she has never seen a zebra [3]. In zero-shot learning, semantic features are needed in addition to using sample features. Semantic features are linguistic descriptions of the samples, such as the color, size and other characteristics. Word vectors extracted by Word2Vec [4] are usually used as semantic features. In zero-shot learning, the model is trained by the seen class samples with semantic features to find the relationship between semantic features and the seen class samples, and then transfers the model to unseen class samples to infer the categories of the unseen class samples.

There are two categories of zero-shot learning: generalized zero-shot learning and conventional zero-shot learning. The test set contains only the unseen class samples for conventional zero-shot learning. While for generalized zero-shot learning, the test set contains not only the unseen class samples but also the seen class samples. In zero-shot learning, there are many approaches devoted to finding the relationship between the training set samples and the training set semantic features, such as mapping the training set samples to the semantic feature space, mapping the semantic features to the sample space [5], mapping the semantic features and samples to the common space [6,7] and mapping the semantic features and samples to each other's space [8]. However, the training set contains only the seen class samples, and the classes in the training set are not the same as those in the unseen class, which can lead to the inaccurate classification of the unseen class samples when the classification models are used in the test set.

To address these problems, some scholars use generative models, such as Variational Autoencoder (VAE) [9] and Generative Adversarial Network (GAN) [10], to generate the unseen class pseudo samples, and input them to the classifier for training, which can alleviate the problem of inaccurate classification of samples in the unseen classes. It has been proposed in the literature [11,12] that the use of VAE and GAN leads to the problems of posterior collapse and training instability, and the use of Wasserstein Auto-Encoder (WAE) and Wasserstein Generative Adversarial Network (WGAN) to generate pseudo samples can alleviate this problem. However, the unseen class samples generated by these methods are easily biased to the features of the seen class samples, leading to inaccurate classification results when classifying the real unseen class samples. To address these problems, we propose the following methods:

1) Different from the above generation models, we use autoencoder to generate the unseen class samples. To make the sample features in the latent space more distinguishable and representative, a classifier is used for the sample features in the latent space.

2) In order to reduce the unseen class pseudo sample features that are biased to the seen class sample features, we propose new sample features and use them together with unseen class semantic features for cross-reconstruction loss function.

3) The proposed method is validated for three datasets, AWA1, AWA2 and aPY and have good results.

The structure of this paper is organized as follows. First, we thoroughly review the related works in Section 2. The proposed method is illustrated in Section 3. In Section 4, we discuss the experiments, and we conclude the paper in Section 5.

2. Related works

For zero-shot learning, embedding-based zero-shot learning is a common approach. However, embedding-based zero-shot learning can produce domain shift problems and misclassification in unseen class samples [13].

There are three solutions to alleviate the unseen class samples that are easily misclassified into the seen classes in zero-shot learning: calibrated stacking, generative models and detection of the unseen class samples. Calibrated stacking [14] added a calibrated term to the classifier so that the score of the seen class is reduced during classification, and the score of the unseen class samples can be increased. The calibrated stacking equation is as follows:

$$\hat{y} = \underset{c \in \mathcal{T}}{\operatorname{argmax}} f_c(x) - \gamma \mathbb{I}[c \in S]$$

where γ is the calibrated factor and the indicator function. $\mathbb{I}[\cdot]$ indicates whether c belongs to a seen class, if c is a seen class, the value of the indicator function is 1, otherwise the value of the indicator function is 0. The classifier of APN [15] used class embedding and added calibrated stacking in the classifier to alleviate misclassification of the unseen class samples.

The method of generative models is to use generative models to generate pseudo samples substitute real unseen class samples [16]. The training set and pseudo samples are used in training the classifier, so that the unseen class samples can avoid bias to the seen classes. Multi-modal Feature Fusion algorithm (MFF) [17] used visual principal component features to compensate for the lack of descriptive information using only semantic features, and then combined GAN and VAE to generate high-quality pseudo-samples. Cross- and Distribution Aligned VAE (CADA-VAE) [18] was the VAE method. The latent space distributional alignment and cross-alignment were used to ensure the alignment between the two different modalities of sample features and semantic features. To make the generated samples close to the real samples, Over-Complete Distribution using Conditional Variational Autoencoder (OCD-CVAE) [19] used over-complete distribution to generate pseudo samples. Chen et al. [12] proposed to use WAE to generate pseudo samples and used an aggregated posterior distribution in the latent space to align the manifold structure of the sample features and the semantic features. f-CLSWGAN [20] used WGAN to generate pseudo samples and used a classifier to make the generated pseudo samples more discriminative. Based on f-CLSWGAN, Adaptive Bias-Aware GAN (ABA-GAN) [21] proposed adaptive adversarial loss and domain loss functions to make the generated pseudo samples more meaningful and to distinguish the seen classes from the unseen classes. Li et al. [11] used WGAN to generate pseudo samples and used multimodal cyclic loss function and bi-directional autoencoder. In response to the fact that GAN is not easy to train, and the pseudo samples generated by VAE are of low quality, Dual VAEGAN [22] used a combination of GAN and VAE.

Detection of samples of the unseen class. This method first distinguishes whether the samples belong to the seen classes or the unseen classes and then classifies the samples into specific class. GatingAE [3] first used the latent space and the cross-reconstruction space to detect samples belonging to the unseen class, and then used a linear classifier to classify the samples in the seen classes and a nearest neighbor classifier for the samples belonging to the unseen classes. Chen et al. [23] proposed determining whether a sample belongs to the seen class or the unseen class by calculating the cosine similarity between the latent space features of the samples and the mean value of each class. However, the models in these methods are obtained by training the samples from the seen classes; when migrating to the samples from the unseen classes, the classification results of the samples from the unseen classes are still biased to the seen classes.

Cao et al. [24] achieved recognition of zero shot traffic signs using autoencoder. Different from the literature [24], we use autoencoder to generate the unseen samples to alleviate the misclassification of the unseen class samples. To prevent the generated unseen class samples biased towards the features of seen class samples and improve the classification accuracy of the unseen class samples, we add the information of both unseen class semantic features and the proposed sample features.

3. The proposed method

3.1. Definition of zero-shot classification

In zero-shot learning, the training set can be denoted as $S = \{X_S, A_S, Y_S\}$,and the unseen class can be denoted as $U = \{X_U, A_U, Y_U\}$, where X denotes sample features, A denotes semantic features and Y denotes labels. For conventional zero-shot learning, the class of X_U is predicted by the classifier: $X_U \rightarrow Y_U$; for generalized zero-shot learning, the class of X is predicted by the classifier: $X \rightarrow Y_S \cup Y_U$.

3.2. Zero-shot learning via visual-semantic aligned autoencoder

In this study, we use autoencoder to generate the pseudo samples of unseen classes, and the model is shown in Figure 1. In the Figure 1, E1 and E2 represent the encoder, D1 and D2 represent the decoder. The sample features and semantic features are encoded to obtain the same dimensional latent space features.

According to the autoencoder, for the training set, the generated sample features $\widetilde{X_S}$ and the semantic features of the seen classes $\widetilde{A_S}$ need to approximate the input features X_S and A_S.Assuming that there are m samples, the reconstruction loss function can be written as:

$$L_{recon1} = \frac{1}{m}\sum_{i=1}^{m}|x_{si} - \widetilde{x_{si}}| + \frac{1}{m}\sum_{i=1}^{m}|a_{si} - \widetilde{a_{si}}| \tag{1}$$

We use the lowercase x_{si}, $\widetilde{x_{si}}$, a_{si} and a_{si} to denote one sample feature in X_S, one generate sample feature in $\widetilde{X_S}$, one semantic feature in A_S and one generated semantic feature in $\widetilde{A_S}$ respectively. We want these two modality features to be aligned in the latent space. We use Z_S to represent the sample features of the latent space and Z_{AS} to represent the seen class semantic features of the latent space.

$$L_{latent-recon} = \frac{1}{m}\sum_{i=1}^{m}|z_{Si} - z_{ASi}| \tag{2}$$

In Eq (2), we use the lowercase z_{Si} to denote one sample feature in Z_S, and use the z_{ASi} to denote one seen class semantic feature in Z_{AS}. In addition to the reconstruction loss function shown in Eq (1), zero-shot learning contains two different modalities, sample features and semantic features. Aligning different modalities can reduce the domain shift problem [25]. Inspired by GatingAE [3], Chen et al. [12], CADA-VAE [18] and Discriminative Cross-Aligned Variational Autoencoder(DCA-VAE) [25] ,we use the cross-reconstruction loss function. The features $\overline{X_S}$ and $\overline{A_S}$ are obtained by passing the semantic features and sample features of the latent space through D2 and D1 decoders, respectively, the cross-reconstruction loss function is as follows:

$$L_{cross-recon1} = \frac{1}{m}\sum_{i=1}^{m}|x_{si} - \overline{x_{si}}| + \frac{1}{m}\sum_{i=1}^{m}|a_{si} - \overline{a_{si}}| \tag{3}$$

Here, $\overline{x_{s\imath}}$ denotes one feature in $\overline{X_S}$, and $\overline{a_{s\imath}}$ denotes one feature in $\overline{A_S}$. Although we can use Eqs (1), (2) and (3) to train the model and then to generate samples of the unseen classes, Eqs (1), (2) and (3) only contains samples of the seen classes and semantic features, which will lead to the pseudo samples being biased to the seen classes. To address this problem, we add the unseen class semantic features A_{US} to the model and propose new sample features $\hat{X}$.

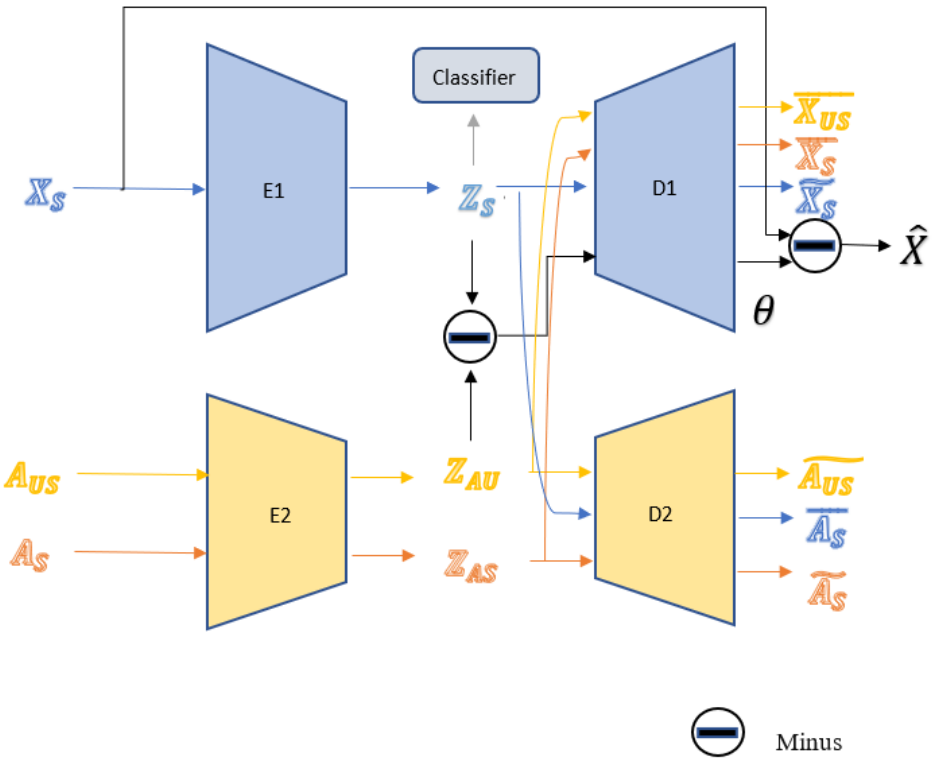

Figure 1. The model of the proposed method.

The unseen class semantic features A_{US} can be obtained by the following method. The sample features of the training set are mapped to the semantic feature space using the following equation:

$$\min_{W} \|X_S - W^T A_S\|_F^2 + \alpha \|W\|_F^2 \tag{4}$$

$\|\cdot\|_F$ in Eq (4) denotes Frobenius norm. The mapping matrix W is obtained as follows:

$$W = X_S^T A_S (A_S^T A_S + \alpha I)^{-1} \tag{5}$$

where I in Eq (5) represents the unit matrix and α denotes an adjustable parameter. The sample features in the training set are then mapped to the semantic feature space through the mapping matrix W and find the nearest unseen class semantic features, which constitute A_{US}.

After obtaining A_{US},we input A_{US} to the autoencoder to obtain the generated unseen class semantic features $\widetilde{A_{US}}$, and the reconstruction loss between $\widetilde{A_{US}}$ and A_{US} is:

$$L_{recon2} = \frac{1}{m}\sum_{i=1}^{m} |a_{usi} - \widetilde{a_{us\imath}}| \tag{6}$$

Here, we use a_{usi} to denote one unseen class semantic feature in A_{US} and $\widetilde{a_{us\iota}}$ to denote one generated unseen class semantic feature in $\widetilde{A_{US}}$. Except the reconstruction loss for A_{US}. We also want to align different modalities between the unseen class semantic features and sample features, but there is a lack of unseen class samples in the training set. In this paper, we take the following approach to get the cross-reconstruction loss function: Find the difference between the unseen class semantic features and the seen class sample features in the latent space, the difference represents the relationship between the unseen class semantic features and the seen class sample features in the latent space. Then, pass the difference through the decoder D1 to get θ. We use Z_S and Z_{AU} to represent the latent features of the seen class sample features and the latent features of the unseen class semantic features. The formula is as follows:

$$\theta = D1(Z_S - Z_{AU}) \tag{7}$$

Then subtract θ from the sample features of the training set to obtain the feature $\hat{X}$:

$$\hat{X} = X_S - \theta \tag{8}$$

The cross-reconstruction loss function for unseen class semantic features can be written as:

$$L_{cross-recon2} = \frac{1}{m}\sum_{i=1}^{m}|\widehat{x_\iota} - \overline{x_{us\iota}}| + \beta\frac{1}{m}\sum_{i=1}^{m}|a_{usi} - \overline{a_{s\iota}}| \tag{9}$$

β in the above equation is an adjustable parameter. $\widehat{x_\iota}$ denotes one feature in $\hat{X}$, $\overline{x_{usi}}$ denotes one feature obtained by passing one unseen class semantic feature through the decoder D1. The reason for using the feature $\hat{X}$ instead of X_S is that $\hat{X}$ can reduce the information of the seen class samples in the loss function, which can alleviate the similarity between the unseen class pseudo samples and the seen class samples.

To better find the relationship between the semantic features of unseen classes and the training set samples in Eq (7), and also make the sample features in the latent space distinguishable and representative, the sample features of the latent space are classified using the cross-entropy loss function:

$$L_{classifier} = -\sum_{i=1}^{m} y_{si}\log \widetilde{y_{s\iota}} \tag{10}$$

$\widetilde{y_{s\iota}}$ in Eq (10) is the predicted label and y_{si} is the true label of the sample features of the latent space.

Combining Eqs (1), (2), (3), (6), (9) and (10), the objective function is:

$$L = L_{recon1} + L_{latent-recon} + L_{recon2} + L_{cross-recon1} + L_{cross-recon2} + L_{classifier} \tag{11}$$

3.3. Zero-shot classification

After the model is trained according to Eq (11), the samples are generated with the sample features X_S and semantic features A_{US}. For generalized zero-shot classification, all the generated seen class samples and unseen class samples need to be input to the classifier for training. For conventional zero-shot classification, only the generated unseen class samples need to be input to the classifier for training.

4. Experiments

4.1. Datasets and parameter settings

Three datasets, AWA1, AWA2 and aPY, are used in our study.

1) AWA1 [26]: The seen class contains 40 categories, and the unseen class contains 10 categories. The number of samples in the seen class is 19832, the number of samples in the unseen class is 5685 and the dimension of the semantic features is 85.

2) AWA2 [27]: The seen class contains 40 categories, and the unseen class contains 10 categories. The number of samples in the seen class is 23527, the number of samples in the unseen class is 7913 and the dimension of the semantic features is 85.

3) aPY [28]: The seen class contains 20 categories, and the unseen class contains 12 categories. The number of samples in the seen class is 5932, the number of samples in the unseen class is 7924, and the dimension of the semantic features is 64.

The sample features and semantic features used in our study are taken from the literature [27]. Following the literature [12], the input dimension of encoder E1 is 2048 dimensions, the output of the first layer is 512 dimensions, and the dimension of the latent space is 128; the dimension of the output of the first layer of encoder E2 is 128.The dimension of the output of the first layer of decoder D1 is 256 and the dimension of output is 2048; the dimension of output of the first layer of decoder D2 is 256. We use the Adam algorithm for optimization, the learning rate is 0.001 and the batch size is 256.

4.2. The results of zero-shot classification

We use the evaluation criteria proposed in the literature [27]. For the conventional zero-shot classification, only the accuracy of classification needs to be calculated:

$$acc = \frac{1}{|C|}\sum_{i}^{|C|}\frac{\#\, correct\ predictions\ in\ i}{samples\ in\ i}$$

For generalized zero-shot classification, not only the classification accuracy of the seen class and the unseen class should be calculated, but also the harmonic mean. Assuming that the classification accuracy of the samples of seen classes is denoted as acc_{tr} and the classification accuracy of the samples of unseen classes is denoted as acc_{ts}, the harmonic mean can be written as:

$$H = \frac{2 \times acc_{tr} \times acc_{ts}}{acc_{tr} + acc_{ts}}$$

The generalized zero-shot classification results and the conventional zero-shot classification results are shown in Tables 1 and 2, where the results of Semantic Autoencoder (SAE) [8], Direct Attribute Prediction (DAP) [26], Indirect Attribute Prediction (IAP) [26] and Structured Joint Embedding (SJE) [29] are from the literature [27]. In Table 1, “ts” represents the classification results of unseen classes and “tr” represents the classification results of the seen classes. From Table 1, the proposed method is 1% less than CADA-VAE [18] for the AWA1 dataset. For the AWA2 dataset, the proposed method is 0.5% better than Chen et al. [23]. For the aPY dataset, the proposed method is 3.1%

higher than DAP [26], while it is 4.9% higher than the generative model Chen et al. [12]. The accuracy of the proposed method on unseen class is higher than other methods.

Table 1. The results of generalized zero-shot learning.

	AWA1			AWA2			aPY		
	ts	tr	H	ts	tr	H	ts	tr	H
SAE [8]	1.8	77.1	3.5	1.1	82.2	2.2	0.4	80.9	0.9
DAP [26]	46.5	68.5	55.4	43.7	70.2	53.3	27.6	55.8	37.0
IAP [26]	2.1	78.2	4.1	0.9	87.6	1.8	5.7	65.6	10.4
SJE [29]	11.3	74.6	19.6	8.0	73.9	14.4	3.7	55.7	6.9
Preserving Semantic Relations (PSR) [30]				20.7	73.8	32.3	13.5	51.4	21.4
f-CLSWGAN [20]	57.9	61.4	59.6						
Zhang et al. [31]	20.7	67.9	38.6				16.1	66.9	25.9
Li et al. [11]	54.9	71.7	62.2						
CADA-VAE [18]	57.3	72.8	**64.1**	55.8	75.0	63.9			
Chen et al. [23]	54.7	72.7	62.4	55.6	76.9	64.2			
Chen et al. [12]	54.5	72.8	62.3	55.2	73.5	63.0	26.7	51.5	35.2
The proposed method	62.4	63.9	63.1	60.6	69.5	**64.7**	31.5	55.3	**40.1**

Table 2 shows the conventional zero-shot classification results. For the AWA1 dataset, the proposed method is slightly lower than f-CLSWGAN [20] and Li et al. [11], which used the GAN model. For the aPY dataset, the method in this paper is slightly lower than the method of Zhang et al. [31] and more accurate than the other methods. The accuracy of the method in this paper is higher than the other methods on the AWA2 dataset.

Table 2. The results of conventional zero-shot learning.

	AWA1	AWA2	aPY
SAE [8]	53.0	54.1	8.3
DAP [26]	44.1	46.1	33.8
IAP [26]	35.9	35.9	36.6
SJE [29]	65.6	61.9	32.9
PSR [30]		63.8	38.4
Cross-Class Sample Synthesis (CCSS) [32]	56.3	63.7	35.5
f-CLSWGAN [20]	**69.9**		
Zhang et al. [31]	68.8		**41.3**
Li et al. [11]	**69.9**		
CADA-VAE [18]	58.8	60.3	
Chen et al. [12]	65.2	65.5	32.7
The proposed method	67.1	**66.1**	39.8

4.3. The influence of parameters

The parameters involved in the model are α，β and the dimensionality of the latent space, where we denote the dimensionality of the latent space as d. The effects of taking different values of α，β and d on the generalized zero-shot classification and the conventional zero-shot classification are shown in Figures 2, 3 and 4.

Figure 2 shows the effects of the parameter α on the zero-shot classification results. The parameter α is used to prevent overfitting. Taking the values of α as 0.1, 1, 10 and 100. It can be seen from Figure 2 that the classification results of the aPY dataset are decreasing and then increasing as α keeps increasing. The results of the AWA2 dataset on the conventional zero-shot classification are increasing all the time, while the values of the generalized zero-shot classification are decreasing and then increasing. The results of AWA1 dataset on the conventional zero-shot classification is decreasing and then increasing, and the harmonic mean is always increasing.

The values of β are taken as 0.001, 0.01, 0.1 and 1. β is used to regulate the relationship between the training set samples and the generated unseen class semantic features, and the value of β is taken small because the training set samples are not the real unseen class samples. From Figure 3, the accuracy of conventional zero-shot classification on the aPY dataset is almost unaffected by the value of β, but the value in the generalized zero-shot classification decreases with increasing of β. The classification accuracy of AWA1 and AWA2 on conventional zero-shot classification is also almost unaffected by the value of β, but the harmonic mean value increases and then decreases with increasing of β.

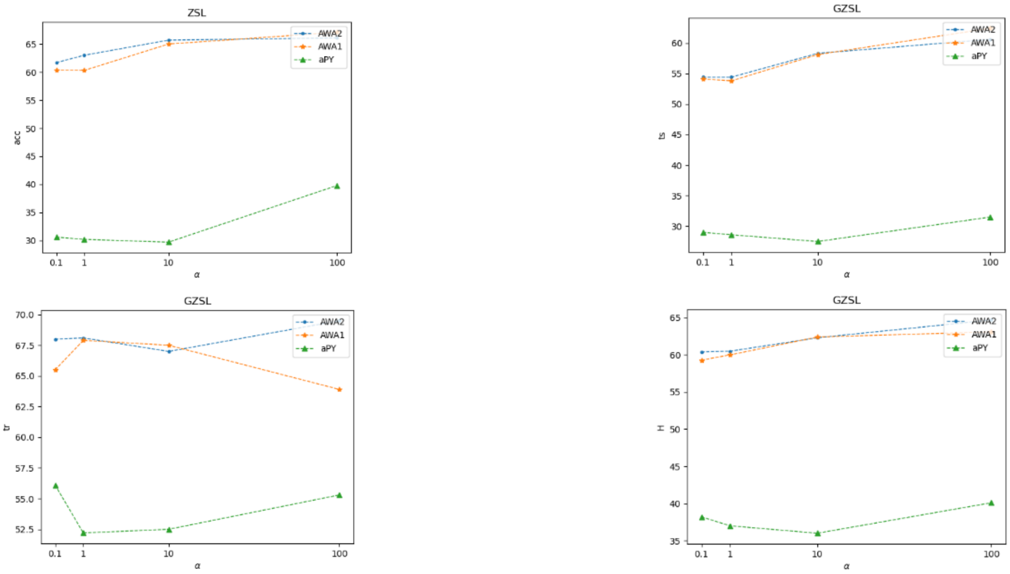

Figure 2. The effects of α on the results of zero-shot classification.

Figure 4 shows the effects of dimension d on the zero-shot classification results, with d taking values of 64, 128 and 256. For the aPY dataset, the accuracy of conventional zero-shot classification increases first and then decreases as d increases, the harmonic mean value keeps decreasing. For the AWA1 dataset, the results increase and then decrease with increasing d, except for the classification results of the seen classes. For the AWA2 dataset, most of the zero-shot classification results show a trend of increasing and then decreasing with increasing d.

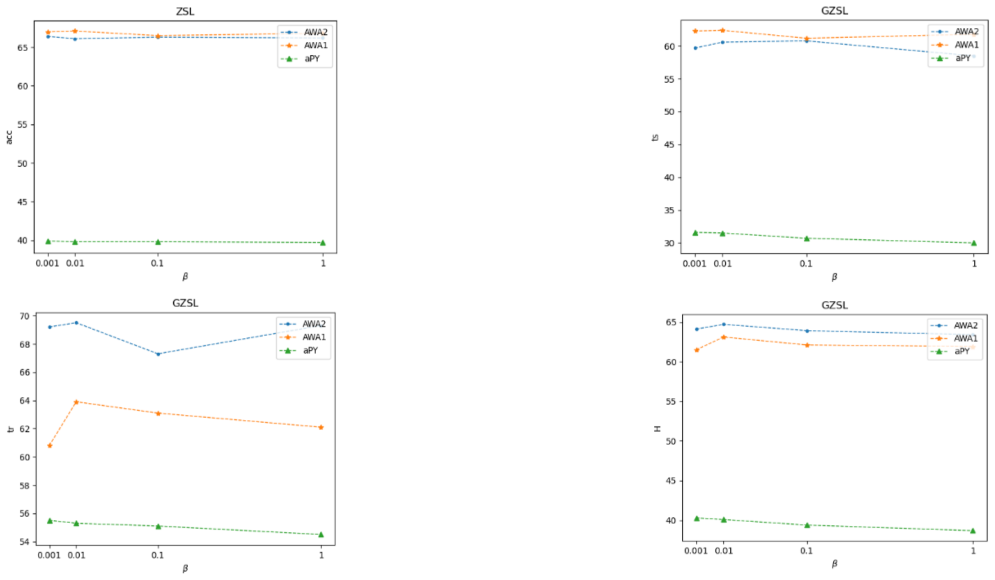

Figure 3. The effects of β on the results of zero-shot classification.

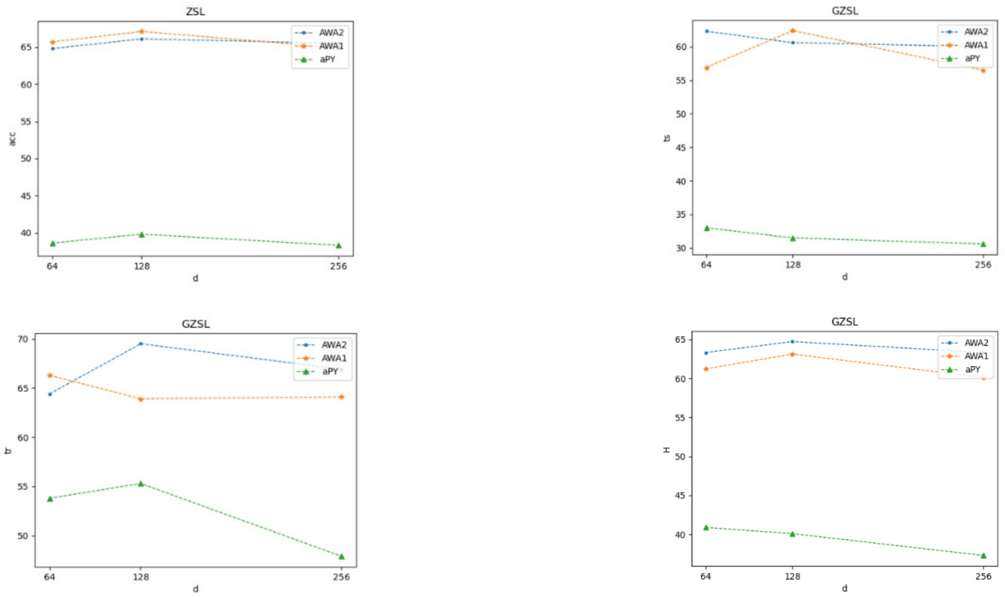

Figure 4. The effects of d on the results of zero-shot classification.

4.4. tSNE

Figure 5 shows the tSNE for generalized zero-shot classification of the aPY dataset, where (a) and (b) denote the training set samples and unseen class samples, respectively, and (c) and (d) are the generated training set samples and unseen class samples.

For the training set samples, the distribution between the generated samples and the original samples is almost the same, and the generated samples are more dispersed between different categories and more concentrated within classes than the original samples. For the unseen class samples, there are more samples presenting orange color in the original samples, while in the generated samples, since A_{US} is chosen for generating the unseen class samples in this paper, it will lead to the number of some classes will be more and the number of some other classes will be less in the generated samples. Except for the inconsistent number of samples, most of the generated samples are similar to the distribution of the real samples.

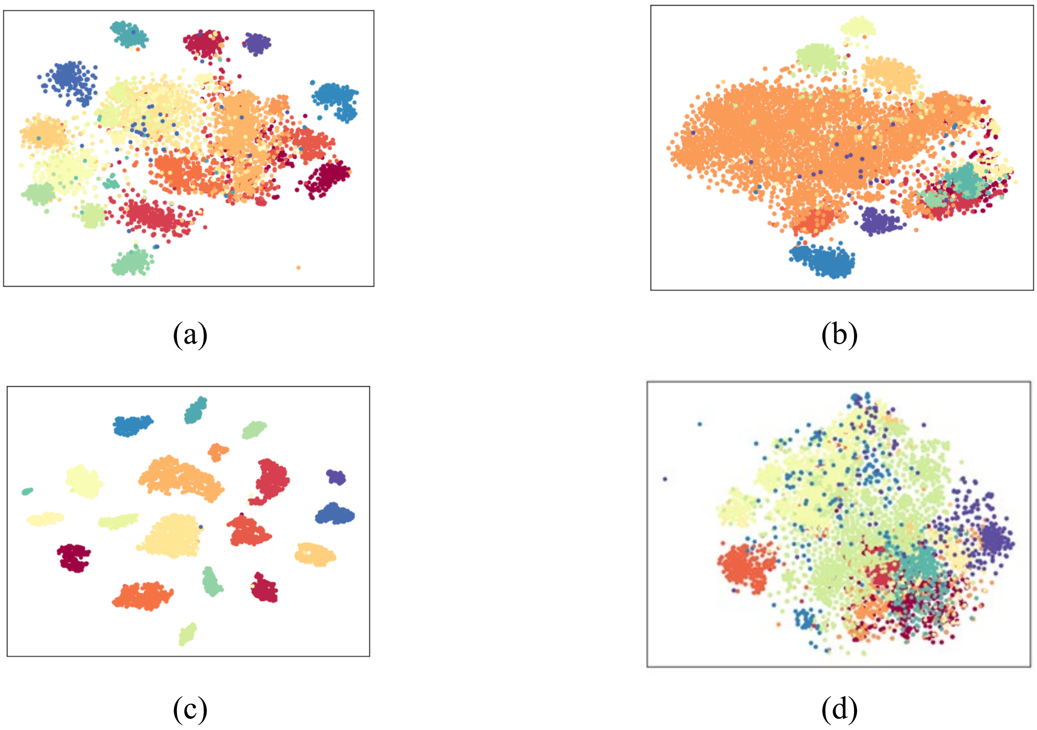
(a) (b) (c) (d)

Figure 5. tSNE of aPY dataset.

4.5. Ablation experiments

The ablation experiments are divided into the following cases: a. Only Eq (1) is retained as the loss function of the model. b. Add $L_{latent-recon}$ to Eq(1). c. $L_{cross-recon1} \neq 0$ on the basis of b. d. $L_{classifier} \neq 0$ on the basis of c; e. Based on d, the second term in $L_{cross-recon2}$ is not 0; f. Based on e, $L_{recon2} \neq 0$. The proposed method is to add the first term in $L_{cross-recon2}$, on the basis of f. The harmonic mean H and the accuracy acc of the conventional zero-shot classification are shown in Table 3.

As can be seen in Table 3, most of the classification results are increased as the term increased in the loss function. However, for the AWA1 dataset when changing from e to f, the accuracy does not change for the conventional zero-shot classification, and the harmonic mean decreases slightly. The proposed method does not increase particularly much in the conventional zero-shot classification compared to other methods, especially in the aPY dataset, and for aPY dataset, the method b is larger than other methods except the proposed method. For AWA2 when changing from b to c, the results decreases, especially in harmonic mean. The seen class information increased when we add the $L_{cross-recon1}$ and the accuracy of the unseen classes decreases. However, for generalized zero-shot classification, the proposed method can provide some information about the unseen classes when training the model, and reduce the similarity between the generated unseen class samples and the seen class samples.

Table 3. The results of ablation experiments.

	AWA1		AWA2		aPY	
	acc	H	acc	H	acc	H
a	45.1	30.0	45.1	17.9	30.7	22.0
b	52.6	32.3	59.4	38.4	38.8	25.0
c	56.5	33.3	57.6	23.4	35.2	26.3
d	59.7	48.3	59.7	40.2	36.8	34.4
e	61.1	51.4	60.2	43.8	36.9	35.1
f	61.1	49.8	61.0	45.7	37.0	36.4
The proposed method	67.1	63.1	66.1	63.1	39.8	40.1

By replacing $\hat{X}$ in the loss function $L_{cross-recon2}$ with X_S, the results are shown in Table 4 numbered as (1), and the results of the proposed model numbered as (2). From Table 4, when X_S is used instead of $\hat{X}$, the results of zero-shot classification are all decreased, especially for generalized zero-shot classification. This is because the loss function contains more information about the samples of the seen classes, making the results easily biased to the seen classes.

Table 4. Comparison between $\hat{X}$ and X_S.

	AWA1				AWA2				aPY			
	acc	ts	tr	H	acc	ts	tr	H	acc	ts	tr	H
(1)	55.1	37.0	71.8	48.8	55.6	34.8	73.0	47.1	35.8	27.0	52.1	35.6
(2)	67.1	62.3	63.9	63.1	66.1	60.6	69.5	64.7	39.8	31.5	55.3	40.1

5. Conclusions

In this study, an autoencoder approach is used for generating samples of unseen classes in zero-shot learning. For the problem that the generated unseen class sample features are always biased to the seen class features, we add the semantic features of the unseen class with the proposed new sample features to the cross-reconstruction loss function. This can reduce the information of the seen class samples and make the generated unseen class samples closer to the real unseen class samples, and improve the classification accuracy of the unseen class samples. The experimental results on three datasets verify that the proposed method can achieve good results.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. R. Gao, X. Hou, J. Qin, Y. Shen, Y. Long, L. Liu, et al., Visual-semantic aligned bidirectional network for zero-shot learning, *IEEE Trans. Multimedia*, **25** (2022), 1649–1664. https://doi.org/10.1109/TMM.2022.3145666
2. L. Yang, X. Gao, Q. Gao, J. Han, L. Shao, Label-activating framework for zero-shot learning, *Neural Netw.*, **121** (2020), 1–9. https://doi.org/10.1016/j.neunet.2019.08.023
3. G. Kwon, G. A. Regib, A gating model for bias calibration in generalized zero-shot learning, *IEEE Trans. Image Process.*, (2022), 1. https://doi.org/10.1109/TIP.2022.3153138
4. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, preprint, arXiv:1301.3781.
5. X. Li, M. Fang, J. Liu, Low-rank embedded orthogonal subspace learning for zero-shot Classification, *J. Visual Commun. Image Representation*, **74** (2021), 102981. https://doi.org/10.1016/j.jvcir.2020.102981
6. Z. Ding, M. Shao, Y. Fu, Low-rank embedded ensemble semantic dictionary for zero-shot learning, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 6005–6013. https://doi.org/10.1109/CVPR.2017.636
7. Y. Liu, X. Gao, J. Han, L. Liu, L. Shao, Zero-shot learning via a specific rank-controlled semantic autoencoder, *Pattern Recognit.*, **122** (2022), 108237. https://doi.org/10.1016/j.patcog.2021.108237
8. E. Kodirov, T. Xiang, S. Gong, Semantic autoencoder for zero-shot learning, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 4447–4456. https://doi.org/10.1109/CVPR.2017.473
9. D. P. Kingma, M. Welling, Auto-encoding variational bayes, preprint, arXiv:1312.6114.
10. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative adversarial nets, in *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, **2** (2014), 2672–2680.
11. J. Li, M. Jing, K. Lu, L. Zhu, H. T. Shen, Investigating the bilateral connections in generative zero-shot learning, *IEEE Trans. Cybern.*, **52** (2022), 8167–8178. https://doi.org/10.1109/TCYB.2021.3050803
12. X. Chen, J. Li, X. Lan, N. Zheng, Generalized zero-shot learning via multi-modal aggregated posterior aligning neural network, *IEEE Trans. Multimedia*, **24** (2022), 177–187. https://doi.org/10.1109/TMM.2020.3047546
13. W. Cao, C. Zhou, Y. Wu, Z. Ming, Z. Xu, J. Zhang, Research progress of zero-shot learning beyond computer vision, in *International Conference on Algorithms and Architectures for Parallel Processing*, **12453** (2020), 538–551. https://doi.org/10.1007/978-3-030-60239-0_36
14. W. Chao, S. Changpinyo, B. Gong, F. Sha, An empirical study and analysis of generalized zero-shot learning for object recognition in the wild, in *European Conference on Computer Vision (ECCV)*, **9906** (2016), 52–68. https://doi.org/10.1007/978-3-319-46475-6_4

15. W. Xu, Y. Xian, J. Wang, B. Schiele, Z. Akata, Attribute prototype network for zero-shot learning, in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS)*, (2020), 21969–21980.
16. W. Cao, Y. Wu, Y. Sun, H. Zhang, J. Ren, D. Gu, et al., A review on multimodal zero-shot learning, *WIREs Data Min. Knowl. Discovery*, **13** (2023), 1488. https://doi.org/10.1002/widm.1488
17. W. Cao, Y. Wu, C. Huang, M. J. A. Patwary, X. Wang, MFF: Multi-modal feature fusion for zero-shot learning, *Neurocomputing*, **510** (2022), 172–180. https://doi.org/10.1016/j.neucom.2022.09.070
18. E. Schönfeld, S. Ebrahimi, S. Sinha, T. Darrell, Z. Akata, Generalized zero- and few-shot learning via aligned variational autoencoders, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 8239–8247. https://doi.org/10.1109/CVPR.2019.00844
19. R. Keshari, R. Singh, M. Vatsa, Generalized zero-shot learning via over-complete distribution, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 13297–13305. https://doi.org/10.1109/CVPR42600.2020.01331
20. Y. Xian, T. Lorenz, B. Schiele, Z. Akata, Feature generating networks for zero-shot learning, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, (2018), 5542–5551. https://doi.org/10.1109/CVPR.2018.00581
21. Y. Yang, X. Zhang, M. Yang, C. Deng, Adaptive bias-aware feature generation for generalized zero-shot learning, *IEEE Trans. Multimedia*, **25** (2023), 280–290. https://doi.org/10.1109/TMM.2021.3125134
22. Y. Luo, X. Wang, F. Pourpanah, Dual VAEGAN: A generative model for generalized zero-shot learning, *Appl. Soft Comput.*, **107** (2021), 107352. https://doi.org/10.1016/J.ASOC.2021.107352
23. X. Chen, X. Lan, F. Sun, N. Zheng, A boundary based out-of-distribution classifier for generalized zero-Shot learning, in *European Conference on Computer Vision (ECCV)*, (2020), 572–588. https://doi.org/10.1007/978-3-030-58586-0_34
24. W. Cao, Y. Wu, C. Chakraborty, D. Li, L. Zhao, S. K. Ghosh, Sustainable and transferable traffic sign recognition for intelligent transportation systems, *IEEE Trans. Intell. Transp. Syst.*, (2022), 1–11. https://doi.org/10.1109/TITS.2022.3215572
25. Y. Liu, X. Gao, J. Han, L. Shao, A discriminative cross-aligned variational autoencoder for Zero-Shot Learning, *IEEE Trans. Cybern.*, **53** (2023), 3794–3805. https://doi.org/10.1109/TCYB.2022.3164142
26. C. H. Lampert, H. Nickisch, S. Harmeling, Attribute-based classification for zero-shot visual object categorization, *IEEE Trans. Pattern Anal. Mach. Intell.*, **36** (2014), 453–465. https://doi.org/10.1109/TPAMI.2013.140
27. Y. Xian, C. H. Lampert, B. Schiele, Z. Akata, Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly, *IEEE Trans. Pattern Anal. Mach. Intell.*, **41** (2019), 2251–2265. https://doi.org/10.1109/TPAMI.2018.2857768
28. A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), 1778–1785. https://doi.org/10.1109/CVPR.2009.5206772
29. Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele, Evaluation of output embeddings for fine-grained image classification, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 2927–2936. https://doi.org/10.1109/CVPR.2015.7298911

30. S. Biswas, Y. Annadani, Preserving semantic relations for Zero-shot learning, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 7603–7612. https://doi.org/10.1109/CVPR.2018.00793
31. H. Zhang, Y. Long, Y. Guan, L. Shao, Triple verification network for generalized zero-shot learning, *IEEE Trans. Image Process.*, 28 (2019), 506–517. https://doi.org/10.1109/TIP.2018.2869696
32. J. Liu, X. Li, G. Yang, Cross-class sample synthesis for zero-shot learning, in *British Machine Vision Conference*, (2018).