IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 13, 2020, accepted May 26, 2020, date of publication May 29, 2020, date of current version June 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2998495

Zero-Shot Classification Based on Word Vector Enhancement and Distance Metric Learning

JI ZHANG[1,2,3,4], YU CHEN[1,2,3,4], AND YONGJIE ZHAI[1], (Member, IEEE)

[1]School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, China
[2]Ubiquitous Power Internet of Things Institute, North China Electric Power University, Baoding 071003, China
[3]Zhenzhong Electric Power Company Limited, Fuzhou 350002, China
[4]Engineering Research Center of Intelligent Computing for Complex Energy Systems, Ministry of Education, North China Electric Power University, Baoding 071003, China

Corresponding author: Yongjie Zhai (zhaiyongjie@ncepu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61773160 and Grant 61871182, and in part by the Natural Science Foundation of Beijing under Grant 4192055.

ABSTRACT The zero-shot classification algorithm has been widely concerned in recent years, in which the labeling of samples of a new category is unnecessary and the cost of annotations can be reduced in applications. This paper presents a zero-shot method for image classification based on word vectors enhancement and distance metric learning. Specifically, the convolutional neural network (CNN) is employed to extract image feature vectors which have the same dimension as semantic feature vectors. Then, an unsupervised learning method is applied on Wikipedia corpus for extracting word vectors and the skip-gram is used to obtain word vectors. The model of analysis dictionary learning is improved by reducing redundant information in word vectors. The obtained sparse vectors are used as semantic features and a distance metric learning method is employed to measure the distance between image features and semantic features. Finally, the classification is implemented by a nearest neighbor based classifier. The effectiveness of the proposed algorithm is validated on the AwA and CUB data sets. Experimental results demonstrate that the proposed method has good performance in terms of both accuracy and robustness.

INDEX TERMS Zero-shot learning, word vectors, analysis dictionary learning, distance metric learning, image classification.

I. INTRODUCTION

Most of the existing object classification methods are within the scope of supervised learning. The accurate identification of certain types of data means that training related models require a large amount of labeled data [1], [2]. However, some categories of data labels are difficult to obtain or require manual labeling of large amounts of data. And the number of object types in the real world continues to be showing a growing trend, which requires the recognition system to continuously increase and reconstruct new data. According to statistics, there are currently about 30,000 types of human identifiable objects [3]. It is arduous to label such a huge amount of data. Therefore, there is an urgent need for a technology that can still identify the data of the target category even if the visual annotation data of the target category is completely missing [4]. Driven by the actual requirement and the continuous development of technology, zero-shot classification technology came into being.

Zero-shot learning (ZSL) in visual classification aims to recognize novel categories for which few or even no training samples are available [5]. Therefore, zero-shot learning can be employed as an effective method to solve the problem of missing class labels [6]. For the current zero-shot classification task, the key to success is to learn the cross-shot mapping relationship between visual and semantic modalities [7]–[9]. In the early stages, most of the zero-shot classification methods can be ascribed to the method of direct attribute prediction(direct attribute prediction, DAP) [10], [11]. To alleviate the unreliability of attribute predictions, Jayaraman *et al.* proposed a novel random forest approach [12], which leverages statistics about each attributes error tendencies in order to select discriminative and predictable decision nodes, thereby obtaining a more robust discriminative model for unseen classes. The model established with this idea has strong interpretability. However, its shortcomings are also evident.

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

For example, the labeling of attributes by humans is not always reliable, and the mislabeled of attributes will have a large negative impact on the performance of such methods. In addition, the correlation between attributes will also lead to the generation of redundant information, which will also seriously affect the performance of the model [13].

In order to better resolve the problems in the DAP method, a method based on the embedding model [14], [15] has proposed. The core idea of the embedding model method is to simultaneously map all visual features and category labels to a certain space, and then perform zero-shot classification based on the similarity measure [16], [17]. The method proposed by Li *et al.* [18], [19] is to learn a distinguishing category description feature from the representation of visual distribution. A nonlinear mapping model with piecewise linear properties was constructed by Xian *et al.* [20], using a ranking-based loss function for training. Akata *et al.* [21] used a bilinear model to establish the compatibility between visual samples and category attribute descriptions, and used 0-1 loss to learn discriminant. Information between different categories.Yu *et al.* presented a direct push classification method [22] for zero-shot images. This paper proposed a structured joint embedding (SJE) model, which textual image features and semantic features of a common feature space through a mapping matrix. So that the sum of the inner products of the two features is maximized. Xian *et al.* proposed a latent embedding model (LatEm) on this basis and achieved good results [23]. However, these methods are easily affected by the hubness phenomenon [24], and the distances between many features are quite close, which leads to a decrease in performance when using the nearest neighbor classification method for classification.

Due to the limitations of using attribute features, this paper uses word vectors to achieve semantic features. Semantic word vector is a high-dimensional vector representation of entity words obtained through unsupervised learning on large-scale text corpora adopting natural language training models. Each category name is under a unique corresponding semantic word vector, thus providing different distance relationship between categories. However, there exists certain redundant information in these semantic word vectors, which affects the effective expression of distance structure information between categories. To reduce redundant information, this paper uses analysis dictionary learning (ADL) to sparsely encode word vectors [25]. In order to better adapt the overall model to the zero-shot classification problem, this paper improves the basic ADL algorithm and proposes the LC-ADL method, which enhances the operational efficiency and has a positive impact on improving the classification accuracy.

When image features and semantic features are mapped to the same feature space, scientific distance measurement methods can accurately reflect the corresponding relationship between them, which is conducive to improving classification accuracy. Traditional ZSL methods are usually measured by Euclidean distances. Images and semantics belong to different modals, if all dimensions of sample features are still configured in equal importance at this time, the relationship between samples cannot be effectively described. In this case, this paper uses distance metric learning (DML) to measure the distance between the image feature vector and the semantic feature vector, and finally uses the nearest neighbor classifier to classify depending upon the distance. In this paper, an improved Large Margin Nearest Neighbor (LMNN) algorithm is used. LMNN shows advantages in this respect and can alleviate the hubness phenomenon to a certain extent. Experimental results show that introducing the combined DML and LMNN to zero-shot learning can achieve satisfactory results and improve the performance of image classification.

The main contributions in the paper are listed as follows:

1) The analysis dictionary learning method is implemented in sparse representation of word vectors to alleviate redundant information.The objective function of the ADL model is improved, and an error term is added to improve the decisiveness of the model. A LC-ADL model combining with a synthetic linear classifier is proposed. It further reduces noise and errors from word vectors.
2) In the distance measurement module, the LMNN algorithm of DML method is introduced. In order to avoid falling into the local optimal solution when using the gradient descent method, reconstructing the loss function can effectively reduce the error rate and the computational complexity. It has better applicability.
3) A zero-shot image classification method based on word vector enhancement and distance metric learning is proposed, which acquires better performance in accuracy and robustness than several mainstream ZSL methods.

II. RELATED WORK

In this section, we introduce the selection of the basic model. We summarize the notations and variables used in this paper in Table 1.

A. ANALYSIS DICTIONARY LEARNING

Word vectors obtained from unsupervised learning from large-scale text corpora, where each dimension contains some redundant information. It will affect the accuracy of the word vector and the effect of the final classification. Therefore, this paper will use the ADL method to enhance the word vector, and sparsely represent the word vector library initially extracted from the corpus. On the one hand, redundancy between the dimensions of these vectors can be folded up, and the information loss caused by compression may be beneficial. On the other hand, more compact vectors are more efficient to calculate.

Dictionary learning can be regarded as a method of data dimensionality reduction, which is mainly divided into two categories: synthesis dictionary learning (SDL) and analysis dictionary learning (ADL). The idea of SDL is that

TABLE 1. Nomenclature.

Sets and matrices	
X	Seen classes dataset
X'	Unseen classes dataset
Y	Semantic feature vectors set
E	Identity matrix of appropriate order
Ω	Analysis dictionary
A	Sparse coding coefficient matrix
I	Dual synthetic linear classifier
Q	Mapping matrix
Vectors	
x	Seen classes image feature
x'	Unseen classes image feature
y	Semantic feature vector
Other symbols	
K_p	K-nearest prior knowledge
$\alpha, \beta, \gamma, \mu$	Scalar parameter
$\|X\|_0$	l_0 norm of matrix
$\|X\|_F$	Frobenius norm of matrix

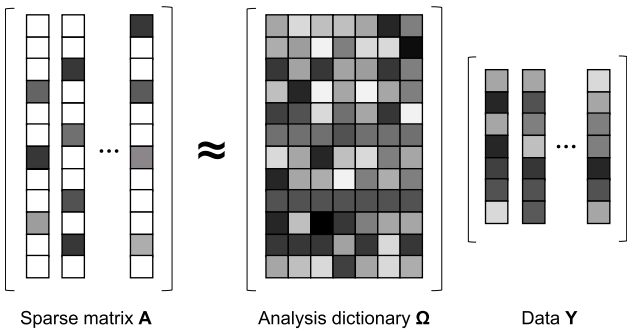


FIGURE 1. The process of sparse coding.

the dictionary and the corresponding sparse coefficients can be reconstructed to obtain the input features. ADL is like a dual structure of SDL. It applies a dictionary to known input features. The sparse coefficients of the input feature can be accessed according to the transformation rules. The advantage of ADL is that its dictionary is obtained by learning related data, which can better adapt to the characteristics of the data. It is highly interpretable and represents the encoding process more intuitively. The schematic diagram of ADL is shown in Figure 1.

In addition, the efficiency of the ADL algorithm for data processing is comparatively high. Given an input feature $\boldsymbol{y} \in \mathfrak{R}^{\boldsymbol{n}}$, the first goal of the ADL algorithm is to learn a parsing dictionary $\boldsymbol{\Omega} \in \mathfrak{R}^{\boldsymbol{m}\times\boldsymbol{n}}$ with a constraint condition of $\|\boldsymbol{a}\|_0 \leq \boldsymbol{T}$. The learned dictionary $\boldsymbol{\Omega}$ satisfies the constraints such that $\|\boldsymbol{a} - \boldsymbol{\Omega y}\|_{\boldsymbol{F}}^{\boldsymbol{2}}$ can achieve a minimum value. The sparseness of sparse coding a is achieved by parameters $\boldsymbol{T}$ and norms l_0. Therefore, the analysis dictionary $\boldsymbol{\Omega}$ can be obtained by solving the following objective function:

$$\begin{aligned} &\min_{\boldsymbol{A},\boldsymbol{\Omega}} \|\boldsymbol{A} - \boldsymbol{\Omega Y}\|_{\boldsymbol{F}}^{\boldsymbol{2}} \\ &\boldsymbol{s.t.}\ \boldsymbol{\Omega} \in \boldsymbol{\Gamma}, \quad \|\boldsymbol{a_i}\|_0 \leq \boldsymbol{T},\ \boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}, \cdots, \boldsymbol{N} \end{aligned} \tag{1}$$

where $\boldsymbol{A} = [\boldsymbol{a_1}, \boldsymbol{a_2}, \cdots, \boldsymbol{a_n}] \in \mathfrak{R}^{\boldsymbol{m}\times\boldsymbol{n}}$ is the sparse coding matrix. Conditions for obtaining a standardized and derivable solution: ① the matrix satisfies the set Γ constraint and the row norm is 1; ② the Frobenius norm that satisfies the matrix is the smallest. The coding coefficient $\boldsymbol{a}$ can be achieved through matrix multiplication and threshold function, which has a high operating efficiency [26], [27].

B. DISTANCE METRIC LEARNING

Traditional measurement methods often use Euclidean distance and cosine distance. However, these methods are not applicable to the case where the importance of each component of the vector is different. After obtaining the image feature vector and the semantic feature vector, a more accurate measurement method is needed to improve the classification effect. Therefore, this paper presents the DML method.

Distance metric learning was proposed by Xing *et al.* [28]. Pan C *et al.* proposed an objective function based on cosine distance to learn the conversion from semantic to visual features [29]. Duan Y *et al.* proposed a deep adversarial metric learning (DAML) framework [30] that can generate synthetic hard negative words from original negative samples. The framework is widely applicable to existing supervised deep metric learning algorithms. In order to take advantage of the nonlinear structure of data points, Hu J *et al.* seek a variety of nonlinear transformations by using neural network architecture [31] and extend MvML to a multi-view deep metric learning (MvDML) method.

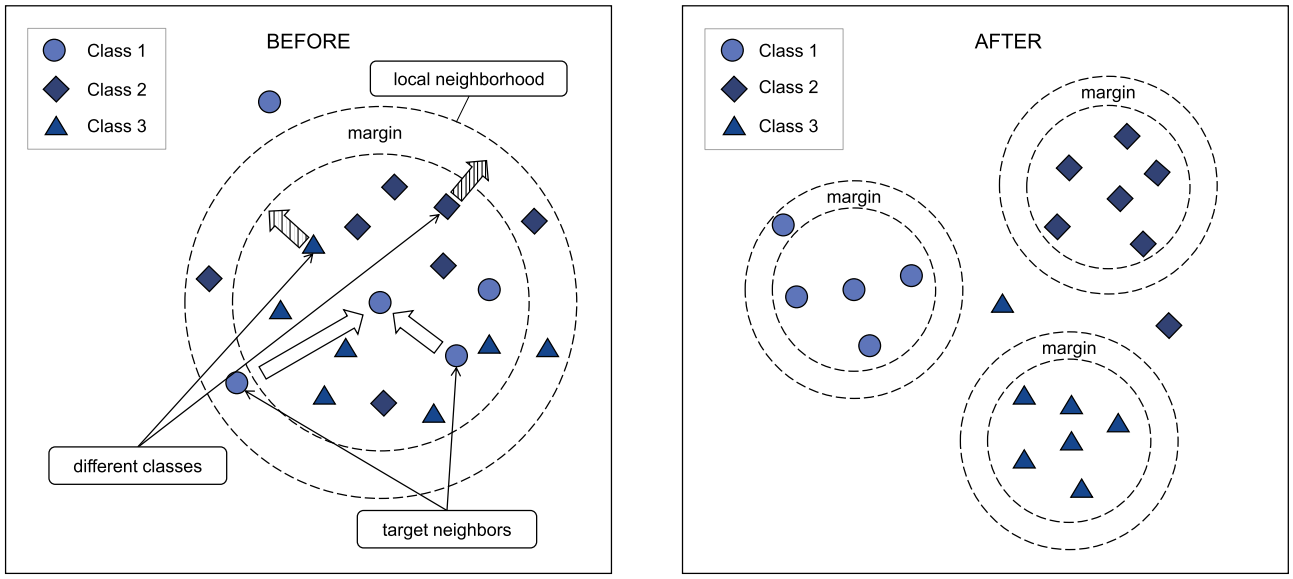
The idea is that for two feature vectors a and b, the learned distance metric form is as follows:

$$\begin{aligned} \boldsymbol{D}(\boldsymbol{a}, \boldsymbol{b}) &= \boldsymbol{D_M}(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\|_{\boldsymbol{M}} \\ &= \sqrt{(\boldsymbol{a} - \boldsymbol{b})^{\boldsymbol{T}}(\boldsymbol{a} - \boldsymbol{b})} \end{aligned} \tag{2}$$

In order to ensure the non-negativity of $\boldsymbol{D_M}(\boldsymbol{a}, \boldsymbol{b})$ and satisfy the triangle inequality, the matrix $\boldsymbol{M}$ should be a semi-positive definite matrix. When $\boldsymbol{M} = \boldsymbol{E}$, $\boldsymbol{D_M}(\boldsymbol{a}, \boldsymbol{b})$ is the Euclidean distance; when $\boldsymbol{M}$ is a diagonal matrix, the elements on the diagonal can be regarded as the weights given to each dimension; when $\boldsymbol{M}$ is a full matrix, the learned distance metric can be counted as the Mahalanobis distance.

C. LARGE MARGIN NEAREST NEIGHBOR ALGORITHM

The goal of DML is to find a metric matrix that minimizes the distance between pairs of similar samples when the sum of the distances between pairs of dissimilar samples is greater than a set fixed value. This paper uses the Large Margin Nearest Neighbor(LMNN) algorithm to cope with this problem. The core of the large interval nearest neighbor algorithm is to replace the Euclidean distance in the traditional K nearest neighbor with the Mahalanobis distance. The LMNN algorithm only penalizes points that are different from the target sample label but are close to it and points that are the same as the target sample label but are far away from it. The k-nearest neighbor prior knowledge of each sample in the training set is necessary for the calculation. The algorithm solves the optimal Mahalanobis distance matrix M through the semi-definite programming optimization method. The optimization maximizes the interval between different classes, so as to ensure that the classification accuracy is improved compared



(a) Before (b) After

FIGURE 2. **The function of the LMNN algorithm.**

with the KNN algorithm. But when the data scale increases, the semi-determined planning scale in the LMNN algorithm also increases greatly, which makes the iteration cost of each step increase, leading to an increase in the computational complexity of the algorithm.

To improve the efficiency of the LMNN algorithm, Shen *et al.* used the gradient descent method to resolve the unconstrained optimization objective function [32]. Weinberger and Saul [33] incorporated slack variables into the objective function, thereby reducing the algorithm complexity. In addition, in order to improve efficiency of the LMNN algorithm, they also proposed a method of spatial mapping using an ellipsoid tree structure. S. Ying *et al.* use the manifold structure of positive-definite matrix group and deduce an intrinsic steepest descent method [34], which assures that the metric matrix is strictly symmetric positive-definite at each iteration, with the manifold structure of the symmetric positive definite matrix manifold. Peng Y *et al.* address the nonlinear metric learning by constructing smooth nonlinear metrics based on data [35]. The partition coefficient obtained by unit partition is smooth, and the metric at any point on the manifold can be directly defined. Huo J *et al.* proposed a CML method by directly maximizing AUC [36]. The method is formulated as a log-determinant regularized semi-definite optimization problem. Li X *et al.* used multiple kernel representation to describe the nonlinear metrics, and projected the data into a high dimensional space where the data can be well represented by linear metric learning [37]. They designed an inherent steepest descent algorithm to learn the positive definite metric matrix.

The function of the LMNN algorithm is illustrated in Figure 2. Figure 2(a) is the original data space. Figure 2(b) is the data space after the LMNN algorithm is mapped. It can be observed that after using the LMNN algorithm, the data of the same category becomes more compact, which is conducive to the accuracy of vector mapping.

The image feature matrix set is $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_s}]$ and the semantic feature set is $\boldsymbol{Y} = [\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_n}]$. n is the sum number of categories. The basic idea of the LMNN algorithm is to set a suitable boundary, learn a training matrix to obtain a mapping matrix $\boldsymbol{L}$, and then mapping the original data with $\boldsymbol{x_i} \rightarrow \boldsymbol{Lx_i}$. The cross-validation method is used on the training set for each point $\boldsymbol{x_i}$ in $\boldsymbol{X}$. It is assumed that $\boldsymbol{x_l}$ in its K-neighbors is different from its class label but within a large margin, and $\boldsymbol{x_j}$ is the same as its class label within a large boundary, the large boundary conditional discriminant can be constructed as follows:

$$\|\boldsymbol{L}(\boldsymbol{x_i} - \boldsymbol{x_l})\|^2 \leq \|\boldsymbol{L}(\boldsymbol{x_i} - \boldsymbol{x_j})\|^2 + \boldsymbol{1} \tag{3}$$

where $\boldsymbol{L}$ is the distance metric matrix, and use this to define non-equivalent constraints. The formula is as follows:

$$\boldsymbol{\varepsilon_{push}}(\boldsymbol{L}) = \sum_{\boldsymbol{i,j} \in \boldsymbol{K_p NN}} \sum\nolimits_{\boldsymbol{l}} (\boldsymbol{1} - \boldsymbol{y_{il}}) \left[\boldsymbol{1} + \boldsymbol{D_L}\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right) - \boldsymbol{D_L}\left(\boldsymbol{x_i}, \boldsymbol{x_l}\right)\right]_+ \tag{4}$$

where K_p is the prior knowledge; and the distance measure of the points $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ after the mapping is: $\boldsymbol{D_L}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \|\boldsymbol{L}(\boldsymbol{x_i} - \boldsymbol{x_j})\|^2$. $j \in K_pNN$ indicates that the training sample $\boldsymbol{x_i}$ is the K-nearest neighbor of the test sample $\boldsymbol{x_j}$; when the semantic vector corresponding to $\boldsymbol{x_i}$ is $y_i = y_l$, $y_{il} = 1$; when $y_i \neq y_l$, $y_{il} = 0$. $[\boldsymbol{Z}]_+ = \boldsymbol{max}(\boldsymbol{Z}, \boldsymbol{0})$. $\varepsilon_{push}(L)$ only affects training samples that are distinct from the test sample category but within the maximum distance. It has a visual effect of 'pushing'.

Similarly, the equivalence constraints formula is as follows:

$$\varepsilon_{push}(L) = \sum_{i,j \in K_pNN} D_L(x_i, x_j) \tag{5}$$

$\varepsilon_{push}(L)$ only affects the training samples that are the same as the test sample category but the distance is beyond the maximum. It has a visual effect of 'pulling'. Finally, combining the formulas (4) and (5) to construct the loss function as follows:

$$\varepsilon(L) = (1-\mu)\varepsilon_{pull}(L) + \mu\varepsilon_{push}(L) \tag{6}$$

where μ is the weight coefficient and we take the value of μ is 0.5. It can be seen that when calculating the mapping matrix $\boldsymbol{L}$, only the points that partially affect the classification by mistake are penalized, which simplifies the computational complexity of obtaining the global optimal mapping and effectively reduce error rates.

III. MODEL IMPROVEMENT

A. LC-ADL MODEL

This paper uses ADL method to sparse the word vector library to achieve the purpose of enhancing the word vector to maximize the useful information. However, the judgment capacity of ADL is not strong and needs to be further improved. We add a classification error term based on a synthetic linear classifier to the objective function of the basic model of ADL. The synthetic linear classifier $\boldsymbol{I}$ uses the dual form of the universal linear classifier: $\boldsymbol{A} \cong \boldsymbol{IC}$. $\boldsymbol{I}$ establishes a corresponding relationship between the coding coefficients and the category labels of the data. The classification error term based on the synthetic linear classifier is:

$$\min_{\boldsymbol{I}} \|\boldsymbol{A} - \boldsymbol{IC}\|_F^2 \tag{7}$$

Therefore, the objective function of the LC-ADL model proposed in this chapter can be optimized as:

$$\begin{aligned} &\min_{\boldsymbol{A},\boldsymbol{\Omega},\boldsymbol{I}} \|\boldsymbol{A} - \boldsymbol{\Omega Y}\|_F^2 + \alpha\|\boldsymbol{A} - \boldsymbol{IC}\|_F^2 \\ &s.t.\ \boldsymbol{\Omega} \in \boldsymbol{\Gamma}, \quad \|\boldsymbol{a}_i\|_0 \le \boldsymbol{T},\ i = 1, 2, \cdots, N \end{aligned} \tag{8}$$

where $\boldsymbol{A} = [\boldsymbol{a_1}, \boldsymbol{a_2}, \cdots, \boldsymbol{a_n}] \in \mathfrak{R}^{m \times N}$ is a sparse coding matrix. $\boldsymbol{C} = [\boldsymbol{c_1}, \boldsymbol{c_2}, \cdots, \boldsymbol{c_m}] \in \mathfrak{R}^{K \times N}$ is the word vector matrix extracted from the Word2Vec model that has been trained on the Wikipedia corpus. K is the sum number of categories of training samples. The function of the parameter α is to control the weight of the classification error term. Γ is the set of the constraint analysis dictionary $\boldsymbol{\Omega}$, and the matrix in the set Γ satisfies the row norm of 1. In addition, in order to ensure that the results reproducible, the matrix of Γ also satisfies the Frobenius norm of the matrix to the minimum. $\boldsymbol{A}$, $\boldsymbol{\Omega}$, and $\boldsymbol{I}$ can be calculated by solving the optimization problem (8). We expand an alternating iterative algorithm to solve the LC-ADL model. The results of formula (8) optimization can be calculated alternately by the following two steps:

1) FIX $\boldsymbol{A}$, UPDATE $\boldsymbol{\Omega}$ AND $\boldsymbol{I}$

According to the constraint set Γ set above, the sub-optimization problem of the analysis dictionary can be described as follows:

$$\boldsymbol{\Omega}^* = \arg\min_{\boldsymbol{\Omega}} \|\boldsymbol{A} - \boldsymbol{\Omega Y}\|_F^2 + \boldsymbol{\beta}\|\boldsymbol{\Omega}\|_F^2 \tag{9}$$

The penalty term $\|\boldsymbol{\Omega}\|_F^2$ in equation (9) is to obtain a stable solution. β is a scalar parameter. After obtaining the optimal solution of formula (9), in order to avoid trivial solutions, each row of $\boldsymbol{\Omega}^*$ must be renormalized to the unit norm. Since the term $\|\boldsymbol{A} - \boldsymbol{IC}\|_F^2$ has no bearing on solving the sub-problems of $\boldsymbol{\Omega}$, this term is omitted in this step. Similarly, the formula for the sub-optimization problem of the classifier is as follows:

$$\boldsymbol{I}^* = \arg\min_{\boldsymbol{I}} \|\boldsymbol{A} - \boldsymbol{IC}\|_F^2 \tag{10}$$

Differentiate the objective function in formula (9) and make its first derivative is equal to 0, and a closed-form solution of $\boldsymbol{\Omega}$ can be obtained:

$$\boldsymbol{\Omega}^* = \boldsymbol{AY}^T(\boldsymbol{YY}^T + \boldsymbol{\beta E})^{-1} \tag{11}$$

Renormalize each line of $\boldsymbol{\Omega}^*$ to the unit norm to get the final solution of the parse dictionary. Similarly, we can get the closed-form solution of $\boldsymbol{I}$:

$$\boldsymbol{I}^* = \boldsymbol{AC}^T(\boldsymbol{CC}^T + \boldsymbol{\gamma E})^{-1} \tag{12}$$

where $\gamma = 10e - 6$ is to ensure that the inverse of $\boldsymbol{CC}^T$ is obtainable. $\boldsymbol{E}$ is the identity matrix corresponding to it.

2) FIX $\boldsymbol{\Omega}$ AND $\boldsymbol{I}$, AND SOLVE $\boldsymbol{A}$

The solution of the coding coefficient $\boldsymbol{A}$ can be obtained according to formula (8), and the conversion process is as follows:

$$\begin{aligned} \boldsymbol{A}^* &= \arg\min_{\boldsymbol{A}} \|\boldsymbol{A} - \boldsymbol{\Omega Y}\|_F^2 + \boldsymbol{\alpha}\|\boldsymbol{A} - \boldsymbol{IC}\|_F^2 \\ &= \arg\min_{\boldsymbol{A}} \operatorname{tr}|(\boldsymbol{A} - \boldsymbol{\Omega Y})(\boldsymbol{A} - \boldsymbol{\Omega Y})^T \\ &\quad + \boldsymbol{\alpha}\operatorname{tr}\left[(\boldsymbol{A} - \boldsymbol{IC})(\boldsymbol{A} - \boldsymbol{IC})^T\right] \\ &= \arg\min_{\boldsymbol{A}} (1+\boldsymbol{\alpha})\operatorname{tr}\left[\boldsymbol{AA}^T\right] \\ &\quad - 2\operatorname{tr}\left[(\boldsymbol{\Omega Y} + \boldsymbol{\alpha IC})\boldsymbol{A}^T\right] \\ &= \arg\min_{\boldsymbol{A}} \|\boldsymbol{A} - \frac{1}{1+\boldsymbol{\alpha}}(\boldsymbol{\Omega Y} + \boldsymbol{\alpha IC})\|_F^2 \\ &\quad s.t.\ \|\boldsymbol{a}_i\|_0 \le \boldsymbol{T}, \quad i = 1, 2, \cdots, N \end{aligned} \tag{13}$$

The result obtained through this process is the best sparse coefficient matrix $\boldsymbol{A}^*$.

B. IMPROVED LMNN DISTANCE METRIC LEARNING ALGORITHM

When measuring the distance between image feature vectors and semantic feature vectors, the traditional Euclidean distance and cosine distance often cannot effectively describe the mapping relations between them. The scientific distance

measurement method can alleviate the hubness phenomenon, which is conducive to improving the classification accuracy. We use the improved LMNN algorithm in the metric learning module. The linear transformation obtained in formula (6) is non-convex, when using stochastic gradient descent (SGD) algorithm it may fall into a local optimal solution. Given different initial matrices, the final results are different. It is not reproducible for some problems so the applicability needs to be strengthened. By reconstructing the formula (6), it can be converted into a semi-definite programming problem. Define the symmetric positive semidefinite matrix $\boldsymbol{Q} = \boldsymbol{L}^T\boldsymbol{L}$ and use matrix $\boldsymbol{Q}$ instead of matrix $\boldsymbol{L}$. The loss function can be defined as follows:

$$\begin{aligned}\varepsilon(\boldsymbol{M}) = (1-\mu)\sum_{i,j\in K_pNN} D_Q\left(x_i, x_j\right) \\ +\mu\sum_{i,j\in K_pNN}\sum_{l}(1-y_{il}) \\ \left[1+D_Q\left(x_i, x_j\right)_Q - D_Q\left(x_i, x_l\right)\right]_+\end{aligned} \tag{14}$$

where

$$\begin{aligned}&D_Q(x_i, x_j) = (x_i - x_j)^T Q(x_i - x_j) \\ &s.t.\ Q \succcurlyeq 0\end{aligned} \tag{15}$$

In order to facilitate the solution in a larger feasible domain, this paper converts the above equation (14) into a convex program. The non-negative relaxation variable ξ_{ijl} is introduced. The non-zero number of ξ_{ijl} can represent the number of intrusive maximum interval samples in the triple. Construct the following positive semi-definite program:

$$\begin{aligned}&\min\ (1-\mu)\sum_{i,j\in K_pNN}\left(x_i - x_j\right)^T Q\left(x_i - x_j\right) \\ &\quad +\mu\sum_{i,j\in K_pNN,l}(1-y_i)\,\xi_{ijl} \\ &\text{s.t.}\ \ ①\xi_{ijl} \geq 0 \\ &\quad ②\mathbf{Q} \succcurlyeq 0 \\ &\quad ③\ (x_i - x_1)^T Q\,(x_i - x_1) \\ &\quad -\ \left(x_i - x_j\right)^T Q\left(x_i - x_j\right) \geq 1 - \xi_{ijl}\end{aligned} \tag{16}$$

Although there are many constraints for this positive semi-definite program, ξ_{ijl} is very sparse. The reason is that the distribution of most samples is reasonable, and only a relatively small number of samples will invade the fields of other samples, resulting in the loss of hinges, so most of the values are 0. This optimization can be resolved by sub-gradient descent method.

C. ZERO-SHOT CLASSIFICATION MODEL

The flowchart of zero-shot image classification model based on word vector enhancement and distance metric learning is shown in Figure 3. The model structure diagram is shown in Figure 4, which mainly includes the following four steps:

Step 1: Extract the image features of the sample. We use VGGNet-19 convolutional neural network model.

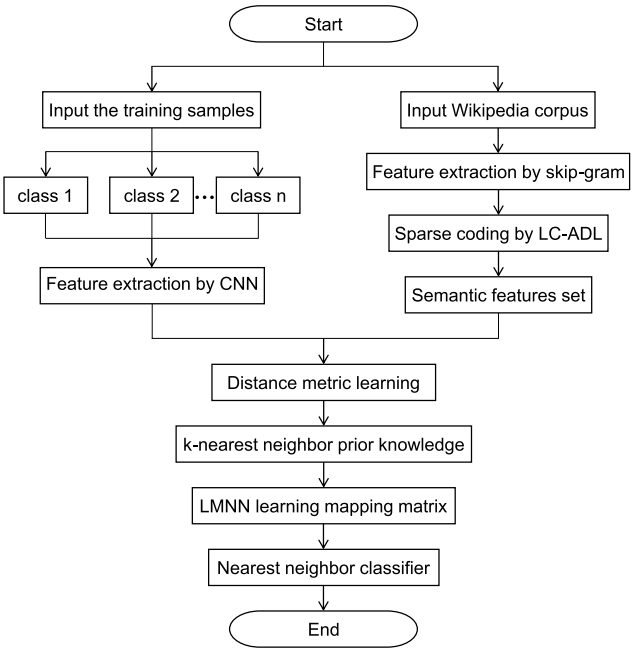


FIGURE 3. The classification process of zero-shot learning model.

As showing in Figure 4, three-channel images of 224 * 224 are input. After convolution and pooling operations, it is finally expanded to generate a 4096-dimensional vector. Add two fully connected layers at the end, and finally output image feature vectors of 200 dimensions.

Step 2: Extract word vectors of all categories. We use skip-gram neural language model for unsupervised learning of large-scale text corpora. Set the dimension of word vectors to 300 dimensions, and each category can get a unique corresponding word vector. As showing in Figure 4, the word vectors of all categories form a category word vector library, whose size is 300 $*$ N. N represents the number of all categories of the sample image. The word vector obtained at this time still contains some redundant information. After LC-ADL processing, the corresponding sparse coding matrix of the word vector library is obtained. Each coding dimension is 200 dimensions.

Step 3: Perform distance metric learning on image feature vectors and semantic feature vectors. Using the Euclidean distance to the training samples, the prior knowledge K-nearest neighbor of each data point in the training set is computed using the cross-validation method, and the label is set. This K value is set to K_p. The improved LMNN algorithm is utilized to learn the mapping rules, and the mapping matrix $\boldsymbol{Q}$ is obtained. The training samples and test samples in the image features are mapped respectively:

$$\begin{aligned}x_i &\rightarrow Qx_i \\ x_i' &\rightarrow Qx_i'\end{aligned} \tag{17}$$

Step 4: Test the sample classification. Utilizing the nearest neighbor classifier, the category corresponding to the text

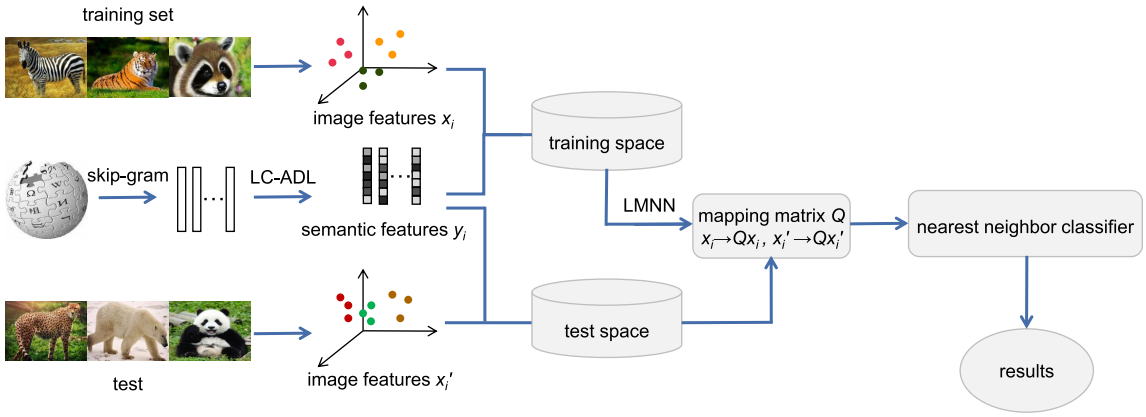


FIGURE 4. The classification process of LC-ADL combined with distance metric learning model.

feature closest to the image feature of the input test sample is the predicted category.

IV. EXPERIMENTS

A. DATA SET AND EXPERIMENTAL SETUP

This paper uses two data sets AwA (animals with attributes) [38] and CUB (caltech-uCSD-birds-200-2011) [39] commonly used in the field of zero-shot learning to verify the proposed model method. Among them, the AwA dataset contains 50 animal categories with a total of 30,475 pictures. The CUB dataset contains 200 bird categories with a total of 11,788 pictures. During the experiment, the default training set test set division method is selected. In AwA, 40 categories are selected as training samples, and the remaining 10 categories are invoked as test samples. In CUB, 150 categories are selected as training samples and the remaining 50 are invoked as test samples. The training samples do not overlap with the test samples. The exact division is shown in Table 2.

In the AwA database, the image features use the same CNN features (VGGNet-19) as in document [40]. Compared to AwA, the CUB dataset is more challenging. Because its objects are birds, the differences between categories are small so it is a data set for fine classification. In addition, the CUB dataset contains more categories, and the number of samples in each category is relatively small, which also increases the difficulty of the CUB dataset. This paper uses the text corpus provided by Wikipedia to extract 300-dimensional semantic features for the category names of AwA and CUB datasets.

B. EVALUATION OF EXPERIMENTAL RESULTS

Since the CUB dataset contains many categories and the time cost of distance metric learning for all samples is very high, this paper randomly samples the training set. 30 samples were selected for each category in AwA; 5 samples were selected for each category in CUB. Using the method described above, random sampling was performed 20 times for repeated experiments. The performance of the algorithm is measured by the average classification accuracy M. Input the images of unseen

TABLE 2. AwA and CUB dataset partition.

AwA		CUB	
Seen classes(40) 24014	Unseen classes(10) 6461	Seen classes(150) 8821	Unseen classes(50) 2967

categories into the model, first classify the classification accuracy within each class, and then calculate the average class accuracy by averaging [41], [42]. The class average accuracy calculation formula is as follows:

$$M = \frac{\sum_{n=1}^{N} Accx_i'}{k}, \quad 1 \le i \le k \tag{18}$$

where k represents the total number of unseen classes, x_i' represents the unseen classes, and $Accx_i'$ is the classification accuracy in the unseen classes.

In order to prove that the combination of LC-ADL and improved DML method can improve classification performance, four groups of experiments will be set up for evaluation:

1) Use Euclidean distance Euc for classification;
2) Use DML method;
3) Use LC-ADL for vector analysis and combined with Euclidean distance;
4) Use LC-ADL for vector analysis and combined with DML;

Table 3 shows the recognition rates of the above four groups of methods performed 20 random trials on the AwA and CUB data sets. It can be observed in the results in Table 3 that the performance of the ADL-DML method has been significantly improved compared with the Euc method. The classification accuracy rate increased by 20.4% in the AwA dataset.In the CUB dataset, it increased by 9.7%. On the one hand, because the semantic feature vectors consist of more noise, the LC-ADL method used in this paper can effectively reduce redundant information and make the semantic vector more accurate. On the other hand, the LMNN algorithm in the

TABLE 3. The accuracy of proposed method and relative methods on AwA and CUB.

Method	AwA(%)	CUB(%)
Euc	46.8±2.0	27.6±1.2
LC-ADL+Euc	51.5±1.6	27.7±0.6
DML	62.3±1.8	30.2±0.8
LC-ADL+DML	**67.2±1.4**	**37.3±0.5**

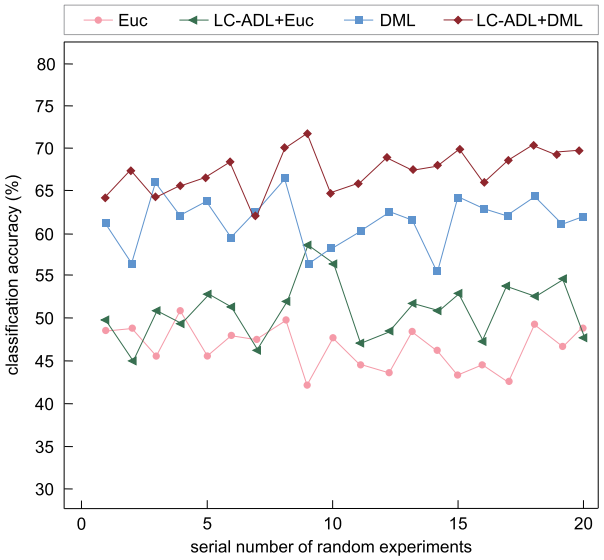


FIGURE 5. The recognition accuracy of the above methods on AwA.

DML method can give lower weight to noise, which has good performance in terms of both noise immunity and robustness. Because the CUB data set is a fine classification, there are many categories and small differences between categories. The LMNN algorithm used in this paper can better handle this problem, so that the elements of the same category are close, and the distance between different categories is farther away, which can alleviate the hubness phenomenon to a certain extent. It is better for classification than Euclidean distance.

The classification accuracy of the four methods on AwA for 20 random times is shown in Figure 5. It can be seen from the figure that our method has better classification effect. The comparison between experiments (1) and (2) proves the effectiveness of the distance metric learning method. The comparison between experiments (1) and (3) shows that sparse coding of the original word vector is beneficial to improve the accuracy.

C. COMPARATIVE ANALYSIS WITH CURRENT MAINSTREAM METHODS

Table 4 displays the average classification accuracy of different algorithms in the AwA and CUB datasets. The comparison algorithms selected in the experiments include DeViSE [43], ESZSL [44], SJE [22], LatEm [23], and Ba et.al [45]. The experimental performance of other comparison algorithms is the value provided by the corresponding article. Figure 6 shows the recognition rate fluctuations of these algorithms using word vectors as semantic features on AwA and CUB.

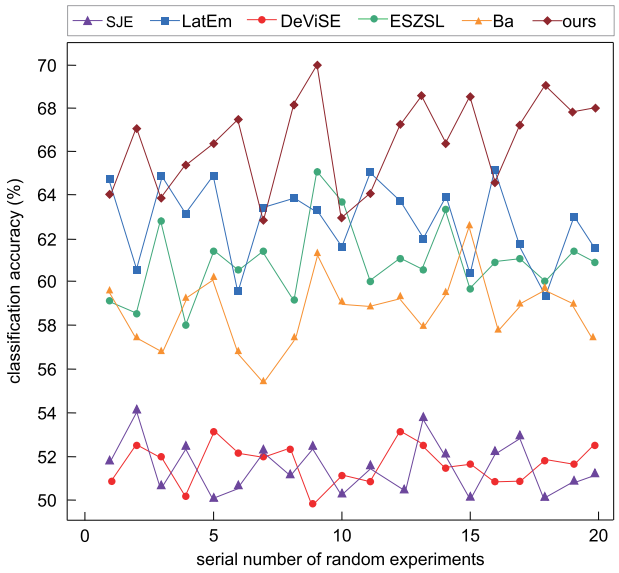


FIGURE 6. Comparison of accuracy with five mainstream methods on AwA.

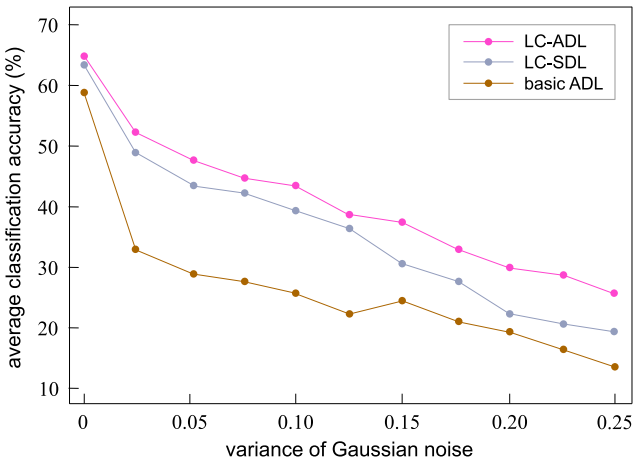
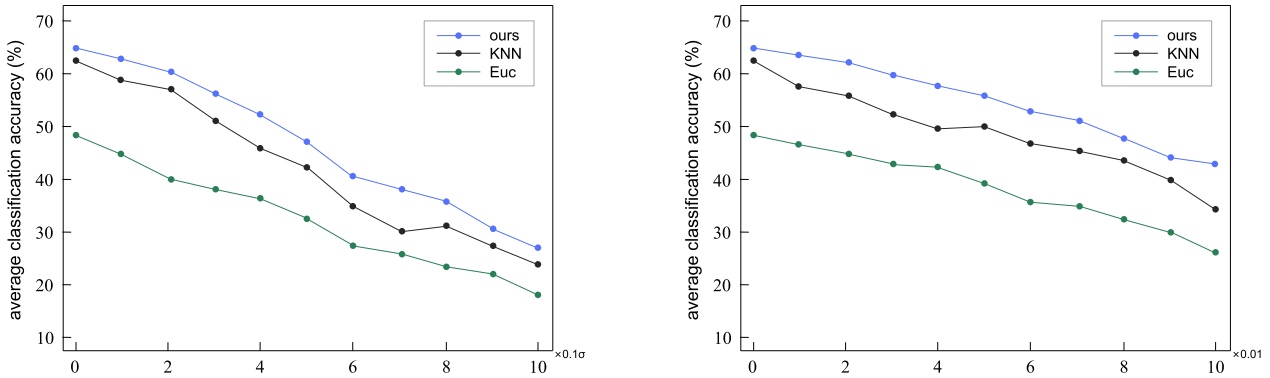


FIGURE 7. The robustness comparison of three algorithms.

It can be seen from Figure 6 that the word vector enhancement and distance metric learning method proposed in this paper can achieve good performance when using word vectors. Referring to Table 4, for the AwA dataset, the performance of this model is 4.9% and 6.6% higher than LatEm and ESZSL. For the CUB dataset, the performance of this model is 3.2% higher than Ba et.al. These results demonstrate the effectiveness of the proposed method of LC-ADL combined with DML.

D. ROBUSTNESS ANALYSIS OF ALGORITHMS

A typical advantage of the dictionary learning method is that it has good robustness for noisy data sets. Therefore, it is necessary to compare the robustness of the LC-ADL algorithm with other algorithms. In order to verify the above points, this paper conducts a comparative experiment. Random Gaussian noise is added to the word vector extracted from the Wikipedia corpus through the skip-gram method,



(a) Gaussian noise (b) Speckle noise

FIGURE 8. Robustness comparison under Gaussian noise and Speckle noise.

TABLE 4. Recognition accuracy of correlation methods on AwA and CUB.

Image feature	Algorithm	AwA(%)	CUB(%)
GoogleNet	SJE [21]	51.2	28.4
	LatEm [22]	62.3	31.8
	DeViSE [33]	52	33.5
	ESZSL [34]	60.6±1.6	32.3±0.8
VGGNet-19	Ba et.al [35]	58.7±1.4	34.1±0.7
	LC-ADL+DML	**67.2±1.4**	**37.3±0.5**

and the variance of the Gaussian noise is gradually increased to verify the robustness of the LC-ADL algorithm.

The comparison algorithm selected in this paper includes the basic ADL method and SDL algorithm. The results of the robustness comparison of the three algorithms are presented in Figure 7. It can be observed in the curve that the LC-ADL algorithm has better performance in robustness than the method based on the synthetic dictionary and the basic ADL method.

In order to examine the robustness of the distance metric learning method, we designed a comparative experiment on the AwA dataset. We add Gaussian noise and Speckle noise to the training images of the AwA dataset. For Gaussian noise, the mean value is 0, and the standard deviation is altered from 0 to 0.1σ, where σ is the standard deviation of the image data. For Speckle noise, we increase its content from 0 to 10%, and observe the changes in algorithm performance.

The comparison algorithm selected in this paper includes KNN method and Euclidean distance. Figure 8 shows the experimental results in these two cases. As can be observed in the curve in the figure, the performance of LMNN is consistently better than KNN and Euclidean distance. It shows that the LMNN method is more robust to noise.

E. EFFECT OF TRAINING SAMPLE NUMBER ON ALGORITHM PERFORMANCE

In practical application, because of the amount of computation and efficiency involved, the training sample is usually used to reduce the training samples. Considering that the number of randomly drawn samples will be related to the

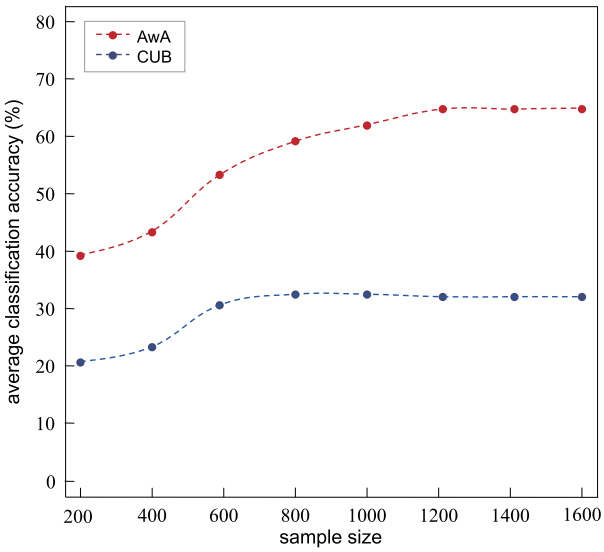


FIGURE 9. Training samples impact on model performance.

experimental results, this paper explores the impact of the number of samples in the training set on the performance of the model. Figure 9 shows the change in ADL-DML performance for different sample numbers in the AwA and CUB datasets.

Figure 9 depicts that as the number of training samples increases, the accuracy rate slowly rises; when the number of samples reaches a certain number, the accuracy rate stabilizes. After comprehensively considering classification performance and calculation amount, the number of samples in AwA and CUB is set to 1200 and 750.

V. DISCUSSION AND CONCLUSION

Based on word vector enhancement and distance metric learning, this paper proposes a zero-shot image classification method, enhancing the accuracy of classification and overcoming the limitation of attribute learning, not necessarily labeling a large amount of data. Word vectors of the corresponding categories are obtained by performing unsupervised

learning on a large amount of text in the Wikipedia corpus. Semantic feature vectors that are more consistent with the distance structure of the image feature vectors can be achieved to improve the robustness of the model by using the improved LC-ADL model to process the word vectors.

We introduce distance metric learning when calculating the correspondence between image feature vectors and semantic feature vectors. LMNN algorithm can effectively alleviate the hubness phenomenon by keeping the elements of the same label within the maximum boundary closer and the elements of different labels far away from each other. The classification results are given by the nearest neighbor classifier according to the distance. When the results are evaluated, in addition to the control factors, the rest of the experiments in different groups use the same network structure and classifier. Based on this, it is proved that the model in this paper has better classification accuracy than the traditional classification model.

When it comes to the Imagenet dataset, one of the limitations of this paper is the increasingly difficult classification due to the growing number of sample categories, including diverse objects, images of animals, plants, objects, scenes, etc., not confined to animals and birds. This part of the task will be placed in our subsequent research work.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[3] I. Biederman, "'Recognition-by-components: A theory of human image understanding': Clarification," *Psychol. Rev.*, vol. 96, no. 1, pp. 115–147, Jan. 1989.

[4] Y. Li and X. Zeng, "Sequential multi-criteria feature selection algorithm based on agent genetic algorithm," *Appl. Intell.*, vol. 33, no. 2, pp. 117–131, 2010.

[5] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks," in *Proc. AAAI Conf. Artif. Intell.*, Chicago, IL, USA, 2008, pp. 646–651.

[6] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2009, pp. 1410–1418.

[7] A. Habibian, T. Mensink, and C. G. M. Snoek, "Video2vec embeddings recognize events when examples are scarce," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 10, pp. 2089–2103, Oct. 2017.

[8] Z. Ji, Y. Yu, Y. Pang, J. Guo, and Z. Zhang, "Manifold regularized cross-modal embedding for zero-shot learning," *Inf. Sci.*, vol. 378, pp. 48–58, Feb. 2017.

[9] D. Elliott, D. Kiela, and A. Lazaridou, "Multimodal learning and reasoning," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, 2016.

[10] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 951–958.

[11] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.

[12] D. Jayaraman and G. Kristen, "Zero-shot recognition with unreliable attributes," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 3464–3472.

[13] D. Jayaraman, F. Sha, and K. Grauman, "Decorrelating semantic visual attributes by resisting the urge to share," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, ON, Canada, Jun. 2014, pp. 1629–1636.

[14] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1060–1069.

[15] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.

[16] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5542–5551.

[17] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal, "A generative adversarial approach for zero-shot learning from noisy texts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1004–1013.

[18] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, and Z. Huang, "Leveraging the invariant side of generative zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7402–7411.

[19] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang, "Designing category-level attributes for discriminative visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 771–778.

[20] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "F-VAEGAN-d2: A feature generating framework for any-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10275–10284.

[21] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2927–2936.

[22] Y. Yu, Z. Ji, X. Li, J. Guo, Z. Zhang, H. Ling, and F. Wu, "Transductive zero-shot learning with a self-training dictionary approach," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 2908–2919, Oct. 2018.

[23] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 69–77.

[24] R. Felix, V. Kumar, I. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 21–37.

[25] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, nos. 1–2, pp. 459–494, Aug. 2014.

[26] M. Yang, H. Chang, and W. Luo, "Discriminative analysis-synthesis dictionary learning for image classification," *Neurocomputing*, vol. 219, pp. 404–411, Jan. 2017.

[27] J. Wang, Y. Guo, J. Guo, M. Li, and X. Kong, "Synthesis linear classifier based analysis dictionary learning for pattern classification," *Neurocomputing*, vol. 238, pp. 103–113, May 2017.

[28] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Proc. Adv. Neural Inf. Process. Syst.* Vancouver, BC, Canada, 2003, pp. 521–528.

[29] C. Pan, J. Huang, J. Hao, and J. Gong, "Towards zero-shot learning generalization via a cosine distance loss," *Neurocomputing*, vol. 381, pp. 167–176, Mar. 2020.

[30] Y. Duan, W. Zheng, X. Lin, J. Lu, and J. Zhou "Deep adversarial metric learning," *IEEE Trans. Image Process.*, vol. 29, no. 1, pp. 2037–2051, 2020.

[31] J. Hu, J. Lu, and Y.-P. Tan, "Sharable and individual multi-view metric learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2281–2288, Sep. 2018.

[32] C. Shen, J. Kim, and L. Wang, "Scalable large-margin Mahalanobis distance metric learning," *IEEE Trans. Neural Netw.*, vol. 21, no. 9, pp. 1524–1530, Aug. 2010.

[33] K. Q. Weinberger and L. K. Saul, "Fast solvers and efficient implementations for distance metric learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1160–1167.

[34] S. Ying, Z. Wen, J. Shi, Y. Peng, J. Peng, and H. Qiao, "Manifold preserving: An intrinsic approach for semisupervised distance metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 2731–2742, Jul. 2018.

[35] Y. Peng *et al.*, in *Proc. SIAM Int. Conf. Data Mining*, 2018, pp. 423–431.

[36] J. Huo *et al.*, *IEEE Trans. Neural Netw. Learn. Syst.*, 2018, pp. 4844–4856.

[37] X. Li *et al.*, *Int. J. Neural Syst.*, vol. 28, no. 2, 2018, Art. no. 1750040.

[38] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.

[39] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. *The Caltech-UCSD Birds-200-2011 Dataset[EB/OL]*. Accessed: Jan. 15, 2011. [Online]. Available: http://www.vision.caltech.edu/visipedia/CUB-200-2011.html

[40] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 2691–2698.
[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015, pp. 1–13.
[42] Z. Fu, T. A. Xiang, E. Kodirov, and S. Gong, "Zero-shot object recognition by semantic manifold distance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 2635–2644.
[43] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "DeViSE: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, 2013, pp. 2121–2129.
[44] B. Romera-Paredes and P. H. S. Torr, "An embarrassingly simple approach to zero-shot learning," in *Proc. 32nd Int. Conf. Mach. Learning.*, Lille, France, 2015, pp. 2152–2161.
[45] J. L. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov, "Predicting deep zero-shot convolutional neural networks using textual descriptions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 4247–4255.

JI ZHANG received the B.Sc. and Ph.D. degrees from North China Electric Power University, China. He is currently an Associate Professor with the School of Control and Computer Engineering, North China Electric Power University. His current research interests include intelligent fault diagnosis, multi-source information fusion, and deep learning.

YU CHEN is currently pursuing the master's degree in computer with North China Electric Power University. Her research interest includes zero-shot learning and image classification based on deep learning.

YONGJIE ZHAI (Member, IEEE) received the B.Sc. and Ph.D. degrees from North China Electric Power University. He is currently a Professor with the Department of Automation, North China Electric Power University. His current research focuses on the application of machine learning and pattern recognition in power systems.