A Simple Discriminative Dual Semantic Auto-encoder for Zero-shot Classification

Yang Liu
State Key Laboratory of ISN,
Xidian University, China
liuyangxidian@gmail.com

Jin Li
Interactive Entertainment Group,
Tencent Inc., China
j.lixjtu@gmail.com

Xinbo Gao
State Key Laboratory of ISN, School of Electronic Engineering,
Xidian University, China
xbgao@mail.xidian.edu.cn

Abstract

Most existing ZSL models focus on searching the mapping between visual space and semantic space directly. However, few models study whether the human-designed semantic information is discriminative enough to recognize different categories. On the other hand, one-way mapping typically suffers from the project domain shift problem. Inspired by the encoder-decoder paradigm, we propose a novel solution to ZSL based on learning a Discriminative Dual Semantic Auto-encoder (DDSA). DDSA aims to build an aligned space to bridge the visual space and the semantic space by learning two bidirectional mappings, which provides us the required discriminative information about the visual and semantic features in the aligned space. The key to the proposed model is that we implicitly exact the principal information from visual and semantic space to construct aligned features, which is not only semantic-preserving but also discriminative. Extensive experiments on five benchmark data sets demonstrate the effectiveness of the proposed approach.

1. Introduction

Most existing ZSL methods pay more attention to directly learn an one-way mapping between the visual and the semantic space, but neglect the function of the reconstruction, which may lead to the domain shift problem [14]. Recently, Liu *et al.* [12] proposed a Graph and Auto-encoder based Feature Extraction (GAFE) model which brings the idea of auto-encoder into ZSL. However, GAFE ignores that whether the human-designed semantic attributes are discriminative enough to recognize different categories. Moreover, the variations within each attribute may be quite large, making it difficult to learn an appropriate classifier. Thus, the learned mapping by GAFE cannot preserve the underlying discriminative information hidden in the data.

To handle the above problems, a Discriminative Dual Semantic Auto-encoder (DDSA) is proposed in this paper. The framework intends to connect the three spaces *i.e.* visual space, aligned space and semantic space together by encoder-decoder paradigm. The established aligned space can remove the irrelevant information in the visual space. Moreover, the discriminative attribute correlations are also implicitly considered in the aligned space.

- The irrelevant information can be removed from the visual space, which is more constructive to establish a reconstruction relationship with the semantic space.
- The aligned attributes can be viewed as the combination of different attributes, thus the aligned space can preserve the semantic information.
- The seen class classifier is utilized to make the aligned attributes discriminative enough to pull the data from the same class together and push those from different classes away from each other.

2. Approach

2.1. Problem Definition

Suppose there are n labeled samples with c seen classes $\{X, S, Y\}$ and n_u unlabeled samples with c_u unseen classes $\{X_u, S_u, Y_u\}$. $X \in R^{d\times n}$ and $X_u \in R^{d\times n_u}$ are d-dimensional visual features in the seen and unseen data, while the corresponding labels are Y and Y_u, respectively. The seen and unseen classes have no label overlap, i.e.,

$Y \cap Y_u = \emptyset$. $S \in R^{k \times n}$ and $S_u \in R^{k \times n_u}$ are k-dimensional semantic representations of instances in the seen and unseen data sets. In the semantic-based classification task, we aim to learn a classifier $f : X_u \rightarrow Y_u$, where the samples in X_u are completely unavailable during training.

2.2. Framework

We propose to learn aligned attributes which can build up the relationship between seen and unseen classes by dual auto-encoders: The first one is *visual space↔aligned space*: learn an auto-encoder between visual space and aligned space. The second one is *aligned space↔semantic space*: learn an auto-encoder between aligned space and semantic space.

We use $A \in R^{m \times n}$ to represent the aligned space. To remove the irrelevant information from the visual space, a linear transformation $W \in R^{m \times d}$ is utilized to build up the relationship between aligned space and semantic space. Then the first auto-encoder can be formulated as:

$$\min_{W,A} \|WX - A\|_F^2 + \left\|W^T A - X\right\|_F^2 \tag{1}$$

To preserve the original semantic information, a linear mapping $Q \in R^{m \times k}$ is utilized to build up the relationship between aligned attributes and original attributes. Thus, the second auto-encoder between aligned space and semantic space aims to solve the following function:

$$\min_{A,Q} \|QS - A\|_F^2 + \left\|Q^T A - S\right\|_F^2 \tag{2}$$

To handle more effective recognition task, the aligned attributes should be discriminative. In other words, we hope to find discriminative attribute combinations to classify different categories. Thus we adopt the classifiers of seen classes to make the aligned attributes more discriminative. Specifically, a mapping $P \in R^{c \times m}$ is learned from the aligned space. In summary, we define the objective function of DDSA as follows:

$$\begin{aligned} \arg \min_{W,Q,P,A} & \|WX - A\|_F^2 + \left\|W^T A - X\right\|_F^2 \\ & + \alpha(\|QS - A\|_F^2 + \left\|Q^T A - S\right\|_F^2) + \beta \|PA - H\|_F^2 \\ s.t. \quad & \|p_i\|_2^2 \leq 1, \quad \forall i, \end{aligned} \tag{3}$$

where $H = [h_1, h_2, \cdots, h_n] \in R^{c \times n}$ and $h_i = [0 \cdots 0, 1, 0 \cdots 0] \in R^c$ is a one-hot vector which represents the class label of seen sample x_i. P can be viewed as an classifier in the aligned space. The last term in Eq. (3) aims to make the aligned attributes discriminative enough to classify different categories.

2.3. Optimization

Obviously, Eq. (3) is not convex for W, Q, P and A simultaneously, but it is convex for each of them separately. Thus, we employ an alternating optimization method to solve the objective function. In particular, we alternate between the following subproblems:

Step 1: Update W while fixing the other variables. The subproblem is formulated as:

$$W^* = \arg\min_W \|WX - A\|_F^2 + \left\|W^T A - X\right\|_F^2. \tag{4}$$

To optimize it, we just need to take a derivative of upper formula and set it to zero. Then we can obtain the formulation as:

$$\left(AA^T\right) W + W \left(XX^T\right) = 2AX^T. \tag{5}$$

Obviously, the Eq. (5) is a Sylvester equation [3] put forward by Bartels and Stewart which can be simply solved by a single line of code in MATLAB[1] .

Step 2: Update Q while fixing other variables. The subproblem is formulated as:

$$Q^* = \arg\min_Q \|QS - A\|_F^2 + \left\|Q^T A - S\right\|_F^2. \tag{6}$$

This problem can be solved in the same way as Eq. (5). The solution for Q is solved by following Sylvester function:

$$\left(AA^T\right) Q + Q \left(SS^T\right) = 2AS^T. \tag{7}$$

Step 3: Update P while fixing other variables. The subproblem is formulated as:

$$\begin{aligned} & P^* = \arg\min_P \|PA - H\|_F^2 \\ & s.t. \quad \|p_i\|_2^2 \leq 1, \quad \forall i. \end{aligned} \tag{8}$$

The above problem can be optimized by the Lagrange dual. Thus the analytical solution for Eq. (8) is:

$$P = \left(HA^T\right) \left(AA^T + \Lambda\right)^{-1}, \tag{9}$$

where Λ is a diagonal matrix constructed by all the Lagrange dual variables.

Step 4: Update A while fixing other variables. The subproblem is formulated as:

$$A^* = \arg\min_A \|M - NA\|_F^2, \tag{10}$$

where

$$M = \begin{bmatrix} WX \\ X \\ \alpha QS \\ \alpha S \\ \beta H \end{bmatrix}, \quad N = \begin{bmatrix} I \\ W^T \\ \alpha I \\ \alpha Q^T \\ \beta P \end{bmatrix}, \tag{11}$$

[1] $W = \text{sylvester}(AA^T, XX^T, 2AX^T)$;

and $I \in R^{m \times m}$ is the m-dimensional identity matrix. Taking a derivative of Eq. (10) and set it zero, we get the closed-form solution for A is:

$$A = \left(N^T N\right)^{-1} N^T M \tag{12}$$

In conclusion, the procedure of optimizing the objective function Eq. (3) is listed in Algorithm 1. In experiments, the optimization process always converges after tens of iterations.

Algorithm 1: DDSA model for zero-shot classification

Input: Data matrix X, semantic matrix S, parameter α and β.
Initialize: Q, P, A randomly
repeat:
1. Update W by solving Sylvester Eq. (5).
2. Update Q by solving Sylvester Eq. (7).
3. Update P by Eq. (9).
4. Update A by Eq. (12).

until converge.
Output: W, Q, P, A.

Verification The classification can be performed in visual space or semantic space. In our experiment, we perform the task in visual space. The learned W and Q project semantic prototypes S_u to the visual space. Then the label of the testing sample X_u^i can be classified by Nearest Neighbour (NN) search with the help of following equation:

$$predict\ label\ (X_u^i) = \arg\min_j d\left(X_u^i, W^T Q S_u^j\right) \tag{13}$$

where X_u^i is the i-th sample of unseen samples. S_u^j is the semantic feature of the j-th unseen class. $d\left(\cdot,\cdot\right)$ represents the Euclidean distance between two vectors.

Table 1: Details of datasets, where s/u means seen/unseen

Dataset	visual dim	s/u classes	s/u samples
SUN	2000	645/72	10320/1440
CUB	2000	150/50	7057/2967
AWA1	2000	40/10	19832/5685
AWA2	2000	40/10	23527/7913
aPY	2000	20/12	5932/7924

3. Experiments

In this section, We validate our proposed specific linear and deep methods on five widely-used data sets and compared with some state-of-the-art models.

3.1. Datasets and Setting

Five widely-used ZSL benchmark data sets are used to verify the effectiveness of the proposed framework. The statistics of all data sets are shown in Table 1.

For a fair comparison, the features and the semantics provided by [20] are used in all experiments. Specifically, the image features are extracted by the 101-layered ResNet [7] and the attribute vectors are utilized as the class semantics. Parameters α and β in our objective function are fine-tuned in the range [0.1,10] using the validation splits. Finally, we set the dimension of the aligned space is 1200, *i.e.* $m = 1200$.

Table 2: ZSL results on SUN, CUB, AWA1, AWA2 and aPY datasets. The results report average per-class Top-1 accuracy in %.

	Method	SUN	CUB	AWA1	AWA2	aPY
Deep	DEVISE [6]	56.5	52.0	54.2	59.7	39.8
	CONSE [15]	38.8	34.3	45.6	44.5	26.9
	CMT [17]	39.9	34.6	39.5	37.9	28.0
	SP-AEN [5]	59.2	55.4	-	58.5	24.1
	PSR [2]	61.4	56.0	-	63.8	38.4
	DCN [11]	61.8	56.2	65.2	-	43.6
	CCSS [10]	56.8	44.1	56.3	63.7	35.5
	fCLSWGAN [21]	58.5	**57.7**	64.1	-	-
Shallow	DAP [9]	39.9	40.0	44.1	46.1	33.8
	IAP [9]	19.4	24.0	35.9	35.9	36.6
	SSE [23]	51.5	43.9	60.1	61.0	34.0
	LATEM [19]	55.3	49.3	55.1	55.8	35.2
	SJE [1]	53.7	53.9	65.6	61.9	32.9
	ESZSL [16]	54.5	53.9	58.2	58.6	38.3
	SYNC [4]	56.3	55.6	54.0	46.6	23.9
	SAE [8]	40.3	33.3	53.0	54.1	8.3
	LESAE [13]	60.0	53.9	66.1	68.4	40.8
	GAFE [12]	62.2	52.6	67.9	67.4	44.3
	DDSA	**63.3**	53.2	**68.3**	**69.1**	**46.1**

3.2. Effectiveness of the Proposed Framework

The proposed aligned space is associated with the visual space and the semantic space. To demonstrate the effectiveness of each component, we compare four different approaches and the ZSL results on SUN data set are shown in Figure 1. (1) learning one auto-encoder between visual space and the aligned space with discriminative constraint (**VD**) (*i.e.* 1, 2, 5 term in Eq. (3)). (2) learning one auto-encoder between semantic space and the aligned space with discriminative constraint (**SD**) (*i.e.* 3, 4, 5 term in Eq. (3)). (3) learning two auto-encoders between visual/semantic space and the aligned space without discriminative constraint (**VS**) (*i.e.* 1, 2, 3, 4 term in Eq. (3)). (3) learning two auto-encoders between visual/semantic space and the aligned space with discriminative constraint (**DDSA**) (*i.e.* Eq. (3)).

By comparing the performance of **VD**, **SD** and **DDSA**, we infer that adopting dual auto-encoders is successful for the ZSL task. Moreover, the performance of **VS** and **DDSA** reflects imposing the discriminative constraint in the objective function can also improve recognition accuracy.

Table 3: GZSL results on SUN, CUB, AWA1, AWA2 and aPY data sets. ts = Top-1 accuracy of the test unseen-class samples, tr = Top-1 accuracy of the test seen-class samples, H = harmonic mean (CMT*: CMT with novelty detection). We measure Top-1 accuracy in %.

		SUN			CUB			AWA1			AWA2			aPY		
	Method	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H
D	DEVISE [6]	16.9	27.4	20.9	23.8	53.0	32.8	13.4	68.7	22.4	17.1	74.7	27.8	4.9	76.9	9.2
e	CMT [17]	8.1	21.8	11.8	7.2	49.8	12.6	0.9	87.6	1.8	0.5	90.0	1.0	1.4	85.2	2.8
e	CMT* [17]	8.7	28.0	13.3	4.7	60.1	8.7	8.4	86.9	15.3	8.7	89.0	15.9	10.9	74.2	19.0
e	CONSE [15]	6.8	39.9	11.6	1.6	72.2	3.1	0.4	88.6	0.8	0.5	**90.6**	1.0	0.0	**91.2**	0.0
p	PSR [2]	20.8	37.2	26.7	24.6	54.3	33.9	-	-	-	20.7	73.8	32.3	13.5	51.4	21.4
	DAP [9]	4.2	25.1	7.2	1.7	67.9	3.3	0.0	**88.7**	0.0	0.0	84.7	0.0	4.8	78.3	9.0
	IAP [9]	1.0	37.8	1.8	0.2	**72.8**	0.4	2.1	78.2	4.1	0.9	87.6	1.8	5.7	65.6	10.4
	SSE [23]	2.1	36.4	4.0	8.5	46.9	14.4	7.0	80.5	12.9	8.1	82.5	14.8	0.2	78.9	0.4
	LATEM [13]	14.7	28.8	19.5	15.2	57.3	24.0	7.3	71.7	13.3	11.5	77.3	20.0	0.1	73.0	0.2
S	SJE [1]	14.7	30.5	19.8	23.5	59.2	33.6	11.3	74.6	19.6	8.0	73.9	14.4	3.7	55.7	6.9
h	ESZSL [16]	11.0	27.9	15.8	12.6	63.8	21.0	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6
a	SYNC [4]	7.9	**43.3**	13.4	11.5	70.9	19.8	8.9	87.3	16.2	10.0	90.5	18.0	7.4	66.3	13.3
l	SAE [8]	8.8	18.0	11.8	7.8	54.0	13.6	1.8	77.1	3.5	1.1	82.2	2.2	0.4	80.9	0.9
l	GFZSL [18]	0.0	39.6	0.0	0.0	45.7	0.0	1.8	80.3	3.5	2.5	80.1	4.8	0.0	83.3	0.0
o	ZSKL [22]	19.8	29.1	23.6	19.9	52.5	28.9	18.3	79.3	29.8	17.6	80.9	29.0	11.9	76.3	20.5
w	LESAE [13]	21.9	34.7	26.9	24.3	53.0	33.3	19.1	70.2	30.0	21.8	70.6	33.3	12.7	56.1	20.1
	GAFE [12]	19.6	31.9	24.3	22.5	52.1	31.4	25.5	76.6	38.2	26.8	78.3	40.0	15.8	68.1	25.7
	DDSA	**22.3**	33.9	**26.9**	**25.1**	53.9	**34.3**	**26.3**	77.1	**39.2**	**28.7**	82.8	**42.6**	**20.4**	62.1	**30.7**

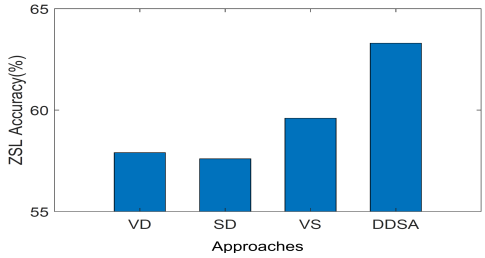

Figure 1: Comparisons of four approaches on SUN data set.

3.3. ZSL and GZSL Results

Comparing the experimental results, we have several interesting observations as follows:

According to Table 2, for ZSL, our model achieves the best results on all data sets except the CUB data set. Specifically, the accuracy of DDSA on the aPY data set increase 4.1% compared the strongest competitor. On other three data sets, the advantage of the DDSA is also obvious. It should be contributed by the learned aligned space. With the help of discriminative aligned attributes, the classification performance of unseen classes can be improved further. Furthermore, CUB is a fine-grained data set where most classes are very similar, so less discriminative structure could be obtained by the DDSA. In contrast, some generative models such as fCLSWGAN or DCN can learn more complicated classifiers to enhance the discriminative property.

According to Table 3, for GZSL, DDSA achieves the highest "ts" and "H" value almost on all data sets, which demonstrates that the discriminative dual auto-encoder structure also benefits the GZSL task. Moreover, it is easy to see that the "ts" value and the "H" value for those baselines with big "tr" value are generally very small. The main reason is that a very big "tr" value reflects the over-fitting training for the seen classes, *i.e.* the trained model in these methods cannot be generalized to new classes.

4. Conclusions

In this paper, we propose a novel ZSL model called Discriminative Dual Semantic Auto-encoder (DDSA). The proposed model aims to learn an aligned attribute space to remove the irrelevant information hidden in the visual space and preserve the semantic information. Furthermore, the aligned attribute space is connected with the similarity space, which makes the aligned attribute space discriminative to recognize different classes. Empirical results on five widely-used data sets show DDSA outperforms existing ZSL models on five benchmarks and the convergence analysis also shows the stability of the proposed algorithm.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grant 61906141, 61432014, 61772402, the National Natural Science Foundation of Shaanxi Province under Grant No. 2020JQ-317, China Postdoctoral Science Foundation (Grant No. 2019M653564), Open Project Program of the State Key Lab of CAD&CG (Grant No. A2018), Zhejiang University and the Fundamental Research Funds for the Central Universities.

References

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015.

[2] Yashas Annadani and Soma Biswas. Preserving semantic relations for zero-shot learning. In *CVPR*, pages 7603–7612, 2018.

[3] Richard H. Bartels and George W Stewart. Solution of the matrix equation ax+ xb= c [f4]. *Communications of the ACM*, 15(9):820–826, 1972.

[4] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, pages 5327–5336, 2016.

[5] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *CVPR*, pages 1043–1052, 2018.

[6] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NeurIPS*, pages 2121–2129, 2013.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[8] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, pages 3174–3183, 2017.

[9] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 36(3):453–465, 2014.

[10] Jinlu Liu, Xirong Li, and Gang Yang. Cross-class sample synthesis for zero-shot learning. In *BMVC*, 2019.

[11] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Generalized learning with deep calibration network. In *NeurIPS*, pages 2009–2019, 2018.

[12] Yang Liu, Quanxue Gao, Jungong Han, Shujian Wang, and Xinbo Gao. Graph and autoencoder based feature extraction for zero-shot learning. In *IJCAI*, pages 15–36, 2019.

[13] Yang Liu, Quanxue Gao, Jin Li, Jungong Han, and Ling Shao. Zero shot learning via low-rank embedded semantic autoencoder. In *IJCAI*, pages 2490–2496, 2018.

[14] Yang Liu, Xinbo Gao, Quanxue Gao, Jungong Han, and Ling Shao. Label-activating framework for zero-shot learning. *Neural Networks*, 121:1–9, 2020.

[15] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[16] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, pages 2152–2161, 2015.

[17] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *NeurIPS*, pages 935–943, 2013.

[18] Vinay Kumar Verma and Piyush Rai. A simple exponential family framework for zero-shot learning. In *ECML-KDD*, pages 792–808, 2017.

[19] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, pages 69–77, 2016.

[20] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, pages 1–1, 2018.

[21] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018.

[22] Hongguang Zhang and Piotr Koniusz. Zero-shot kernel learning. In *CVPR*, pages 7670–7679, 2018.

[23] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, pages 4166–4174, 2015.