Adaptive Metric Learning For Zero-Shot Recognition

Huajie Jiang, *Student Member, IEEE*, Ruiping Wang, *Member, IEEE*, Shiguang Shan, *Senior Member, IEEE*, and Xilin Chen, *Fellow, IEEE*

***Abstract*—Zero-shot learning (ZSL) has enjoyed great popularity in recent years due to its ability to recognize novel objects, where semantic information is exploited to build up relations among different categories. Traditional ZSL approaches usually focus on learning more robust visual-semantic embeddings among seen classes and directly apply them to the unseen classes without considering whether they are suitable. It is well known that domain gap exists between seen and unseen classes. In order to tackle such problem, we propose a novel adaptive metric learning approach to measure the compatibility between visual samples and class semantics, where class similarities are utilized to adapt the visual-semantic embedding to the unseen classes. Extensive experiments on four benchmark ZSL datasets show the effectiveness of the proposed approach.**

***Index Terms*—Zero-shot learning, visual-semantic embedding, adaptive metric learning.**

I. Introduction

OBJECT recognition has made great progress in recent years with the rapid development of deep learning approaches [1]–[4] which require large numbers of images to train robust recognition models [5]. However, collecting and labeling large numbers of images is a difficult task since the number of images for each class follows a long-tail distribution [6]. Therefore, it is difficult to collect images for some uncommon categories. Moreover, traditional supervised learning approaches can only recognize seen-class objects, which is not flexible to novel categories. These challenges motivate the rise of zero-shot learning, where no labeled images are required to recognize a specific category.

Inspired by the human's ability to recognize novel objects, ZSL aims to recognize objects that have never been seen before [7]–[10]. It is achieved by transferring knowledge from seen classes to the unseen classes, where semantic information plays an important role to build up the relations. Traditional ZSL approaches usually involves three steps. First, choose a semantic space to build up the relations among different classes, where the most popular semantic information includes attributes [7], [8], [11]–[14] and texts [15]–[17]. Second, learn general visual-semantic embeddings using seen-class samples. Third, project visual samples and class semantics into a common space and recognize the samples by nearest neighbor approach. Considering different embedding directions, current ZSL approach can be casted into three groups: Some approaches learn transformations from the visual space to the semantic space [7], [11], [18]; Some approaches learn transformations from the semantic space to the visual space [19]–[22]; Other approaches learn a latent space to relate the visual samples and class semantics [15], [23]–[26]. Moreover, recent approaches utilize generative models to generate images features for unseen classes [27]–[29]. [30] leverages hard negative minning strategy to learn more robust visual-semantic embeddings.

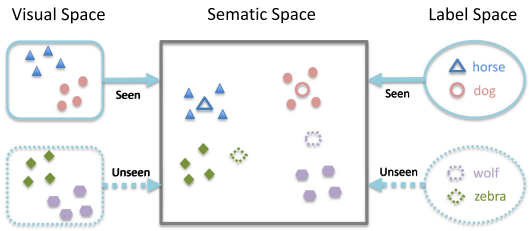

Fig. 1. Illustration diagram for domain shift problem. The embedding features of seen-class samples lie around their class semantics but bias exists for the unseen classes.

Current ZSL approaches usually focus on learning more robust visual-semantic embedding models to make better zero-shot recognition, such as learning multiple embeddings [18] or adding additional regularizations [19]. However, most approaches learn the transformations among seen classes and directly apply them to the unseen classes. It may be not suitable since the domain shift problem exists between seen and unseen classes [31], as is shown in Fig. 1. [31] tries to deal with this problem in transductive settings, where all unseen-class samples are utilized to adjust the model. However, unseen-class samples are often unattainable in real conditions, so we study the inductive ZSL in this letter.

In order to tackle the domain shift problem, we propose a novel adaptive metric learning approach to learn more effective visual-semantic embeddings. The objective of our approach is

Manuscript received March 20, 2019; revised April 30, 2019; accepted May 7, 2019. Date of publication May 15, 2019; date of current version July 24, 2019. This work was supported in part by the 973 Program under Contract 2015CB351802, in part by the Natural Science Foundation of China under Contract 61772500, in part by the Frontier Science Key Research Project CAS QYZDJ-SSWJSC009, and in part by the Youth Innovation Promotion Association CAS 2015085. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shankar Chowdhury. *(Corresponding author: Ruiping Wang.)*

H. Jiang is with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China, with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China, with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: huajie.jiang@vipl.ict.ac.cn).

R. Wang, S. Shan, and X. Chen are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: wangruiping@ict.ac.cn; sgshan@ict.ac.cn; xlchen@ict.ac.cn).

Digital Object Identifier 10.1109/LSP.2019.2917148

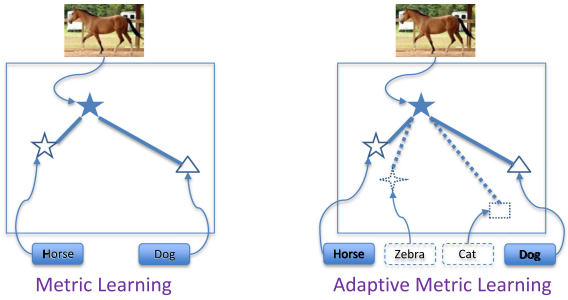

Fig. 2. Comparison between metric learning and adaptive metric learning. The dotted line represents unseen classes.

to learn an embedding function to maximize the compatibility between visual samples and their corresponding class semantics. We use metric learning constraints to ensure that the compatibilities of images with corresponding class semantics are larger than those with different class semantics. Thus we can use the compatibility scores to perform zero-shot recognition task. Since we have no labeled images for unseen classes for training, the model learned on seen classes may have domain shift problem. In order to tackle this problem, we use similar seen-class samples to learn the target classes. Specifically, we add an additional constraint that the compatibilities of seen-class images with similar unseen classes are larger than those with other unseen classes. Thus the model will be more extensive to the unseen classes.

The main contribution of this letter lies in adaptive metric learning, which adjusts the embedding models to the unseen classes using similar seen-class images, thus to make the model more robust to the unseen classes.

II. Adaptive Metric Learning

The framework of our adaptive metric learning (AML) approach is shown in Fig. 2. Different from traditional metric learning approach which enforces the distance of one image with its class semantics is smaller than those with different classes, we adapt such metric learning to the unseen classes by class similarities. Specifically, we constraint that the distance of one image with its similar unseen class is smaller than those with other unseen classes. Thus the learned model will be more robust to the unseen classes.

A. Framework

Assume there are K seen classes (denoted as $\mathcal{Y}$) and L unseen classes (denoted as $\mathcal{Z}$), where seen and unseen classes are disjoint, *i.e.* $\mathcal{Y} \bigcap \mathcal{Z} = \emptyset$. Given N_s labeled images of seen classes $\mathcal{D}_s = \{(x_i, y_i)|x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}_{i=1}^{N_s}$, where x_i represents the image and y_i represents the class label, the goal of ZSL is to learn image classifiers $f_{zsl} : \mathcal{X} \to \mathcal{Z}$ and the goal of generalized zero-shot learning (GZSL) is to learn image classifiers of all classes $f_{gzsl} : \mathcal{X} \to \{\mathcal{Y} \bigcup \mathcal{Z}\}$.

The key to ZSL is how to build up the relations between the visual space and the semantic space. In this letter, we aim to learn a metric to maximize the compatibility between images and their corresponding class semantics:

$$F(x, y; W) = \theta(x)^T W \varphi(y) \tag{1}$$

where $\theta(x)$ represents the image features and $\varphi(y)$ is the class semantic embedding. Then we can use the compatibility scores to perform classification task, as is similarly done in [11].

In order to learn such compatibility functions, we use metric learning constraints to ensure the compatibility of corresponding image and class semantic is larger than those of different ones by a margin Δ:

$$l(x_i, y_i, y) = \max(0, \Delta + F(x_i, y; W) - F(x_i, y_i; W)) \tag{2}$$

where $y \in \mathcal{Y}$ and $y \neq y_i$. This constraint enforces the learned metric to be discriminative enough to separate different classes. The loss function on the seen classes is:

$$L_1 = \frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{y \in \mathcal{Y}, y \neq y_i} l(x_i, y_i, y) \tag{3}$$

In order to enhance the discrimination of the compatibility metric and speed up the learning process, we use the hard negative mining strategy to optimize the model, where the most informative samples are used. The loss function can be formulated by:

$$L_s = \frac{1}{N_s} \sum_{i=1}^{N_s} \max_{y \in \mathcal{Y}, y \neq y_i} l(x_i, y_i, y) \tag{4}$$

It is well known that the domain shift problem exists between seen and unseen classes [31], so the metrics learned on the seen classes may be unsuitable to the unseen classes. Since we have no labeled images for unseen classes, an intuitive idea is to use similar seen-class samples as a substitution. Specifically, for each seen-class image x_i, we choose the most similar class z_i among the unseen classes and make additional metric learning constraints, which enforce the compatibility of one image with similar unseen class is larger than those with other unseen classes by a margin Δ_1. In this way, the seen-class images are fully utilized to transfer their knowledge to the unseen classes and the learned models will be more robust to the unseen classes. Similarly, the loss can be formulated as:

$$l(x_i, z_i, z) = \max(0, \Delta_1 + F(x_i, z; W) - F(x_i, z_i; W)) \tag{5}$$

where $z \in \mathcal{Z}$ and $z \neq z_i$. The loss function on the unseen classes is:

$$L_u = \frac{1}{N_s} \sum_{i=1}^{N_s} \max_{z \in \mathcal{Z}, z \neq z_i} l(x_i, z_i, z) \tag{6}$$

In this way, the metric will be more suitable to the unseen classes. On the whole, the full loss function of the adaptive metric learning approach is:

$$L = L_s + \alpha L_u + \beta L_r \tag{7}$$

where $L_r = ||W||_2$ is the regularization loss.

TABLE I
STATISTICS FOR ATTRIBUTE DATASETS: APY, AWA, CUB AND SUN IN TERMS OF IMAGE NUMBERS (*Img*), ATTRIBUTE NUMBERS (*Attr*), TRAINING + VALIDATION SEEN CLASS NUMBERS (*Seen*) AND UNSEEN CLASS NUMBERS (*Unseen*)

Dataset	*Img*	*Attr*	*Seen*	*Unseen*
APY [8]	15,339	64	15 + 5	12
AWA [7]	30,475	85	27 + 13	10
CUB [33]	11,788	312	100 + 50	50
SUNA [34]	14,340	102	580 + 65	72

B. Class Similarities

In order to accomplish adaptive metric learning, a similar unseen class needs to be selected for each seen-class sample. In this letter, we use the cosine distance to compute the similarities between seen and unseen classes. Specifically, for each seen-class sample x_i, the similar unseen class z_i can be chosen as:

$$z_i = \max_{z \in \mathcal{Z}} \frac{\varphi(y_i)^T \varphi(z)}{||\varphi(y_i)||_2 ||\varphi(z)||_2} \tag{8}$$

Then we can use image x_i and class z_i to adapt the model to the unseen classes.

C. Zero-Shot Recognition

Since the learned metric W measures the compatibility between a visual sample and a class semantic, it can be directly applied to the unseen classes for image classification. Specifically, we can classify one image to the class which has the maximum compatibility score:

$$z = \max_{z \in \mathcal{Z}} \theta(x)^T W \varphi(z) \tag{9}$$

Similarly, we can use such metric to perform GZSL task.

III. EXPERIMENTS

A. Datasets and Settings

We conduct experiments on four widely used ZSL datasets, *i.e.* APY [8], AWA [7], CUB [32], SUN [33], following the newly proposed data splits proposed by [34]. We use AWA1 dataset in this letter. The details of each dataset and class splits for seen and unseen classes are shown in Table I.

To make fair comparisons with other approaches, we use the ResNet101 image features and class attributes provided by [34]. The hyperparameters Δ and Δ_1 are set as 1 and 0 respectively. Other hyperparameters α and β are chosen by cross-validation. We use the limited memory quasi-Newton method L-BFGS [35] to optimize the model by gradient descent.

B. Performance of ZSL

In order to demonstrate the effectiveness of the proposed adaptive metric learning approach, we compare our approach with several ZSL approaches. Table II shows the comparison results, where the performance is evaluated by average per-class top-1 accuracy. The first three approaches utilize nonlinear neural networks to learn the visual-semantic transformations and other approaches are linear embeddings. It can be seen that, among the linear approaches, our approach achieves the best performance on three datasets and is comparable to the best approaches on AWA, which shows that our approach is robust. It is important to point out that the most related approach to ours are [11], [15], which also learn the compatibility function to perform ZSL task. Compared with them, our approach achieves more robust performance on all datasets. The difference mainly lies in the adaptive metric learning constraint. Since we leverage the seen-class images to learn similar unseen classes, the model may be more robust for the unseen classes. The improvement on CUB is more obvious probably due to that CUB is a fine-grained dataset where the seen-class images are more similar to the real unseen-class images and thus the adaptive learning is more effective. Compared with the nonlinear approaches, our approach achieves lower performance since the nonlinear neural networks may be more powerful to learn robust visual-semantic transformations. The idea of adaptive learning is to use seen-class images to learn similar unseen classes, which could also be applied to the nonlinear ZSL models to improve the performance. We will study it in the future.

TABLE II
ZERO-SHOT LEARNING RESULTS ON APY, AWA, CUB AND SUNA UNDER PURE ZSL DATA SPLITS (TOP 1 ACCURACY IN %)

Method	APY	AWA	CUB	SUNA
RNet [25]	-	**68.2**	55.6	-
GAZSL [28]	**41.3**	**68.2**	55.8	**61.3**
FGN [27]	-	**68.2**	**57.3**	60.8
DAP [7]	33.8	44.1	40.0	39.9
IAP [7]	36.6	35.9	24.0	19.4
CONSE [10]	26.9	45.6	34.3	38.8
CMT [9]	28.0	39.5	34.6	39.9
SSE [37]	34.0	60.1	43.9	51.5
LATEM [18]	35.2	55.1	49.3	55.3
DEVISE [23]	39.8	54.2	52.0	56.5
EZSL [19]	38.3	58.2	53.9	54.5
SYNC [20]	23.9	54.0	55.6	56.3
SAE [21]	8.3	53.0	33.3	40.3
ALE [11]	39.7	59.9	54.9	58.1
SJE [15]	32.9	**65.6**	53.9	53.7
AML	**41.6**	65.3	**57.5**	**58.1**

C. Performance of GZSL

Generalized zero-shot learning aims to learn a model to recognize all classes [37]. It not only considers the unseen classes but also the seen classes in the test stage, so it is a more reasonable evaluation of the ZSL models. In order to make fair comparisons with other approaches, we follow the data splits proposed by [34] and evaluate the performance by average per-class top-1 accuracy. The performance of different approaches is shown in Tabel III. It can be seen that most linear approaches achieve very high performance on the seen classes and extremely low performance on the unseen classes. They are prone to overfitting the seen classes. When the recognition scope becomes large, the performance of unseen classes drops greatly, as can be seen by the comparison results of 'ts' and those in Table II. This indicates that most unseen-class samples are recognized as seen classes.

TABLE III
GENERALIZED ZERO-SHOT LEARNING RESULTS ON APY, AWA, CUB AND SUNA. TS = TOP-1 ACCURACY OF THE TEST UNSEEN-CLASS SAMPLES, TR = TOP-1 ACCURACY OF THE TEST SEEN-CLASS SAMPLES, H = HARMONIC MEAN. WE MEASURE TOP-1 ACCURACY IN %

Method	APY			AWA			CUB			SUNA		
	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H
RNet [25]	-	-	-	31.4	**91.3**	46.7	38.1	**61.4**	47.0	-	-	-
GAZSL [28]	**14.2**	**78.6**	**24.0**	29.6	84.2	43.8	31.7	61.3	41.8	22.1	**39.3**	28.3
FGN [27]	-	-	-	**57.9**	61.4	**59.6**	**43.7**	57.7	**49.7**	**42.6**	36.6	**39.4**
DAP [7]	4.8	78.3	9.0	0.0	88.7	0.0	1.7	67.9	3.3	4.2	25.1	7.2
IAP [7]	5.7	65.6	10.4	2.1	78.2	4.1	0.2	**72.8**	0.4	1.0	37.8	1.8
CONSE [10]	0.0	**91.2**	0.0	0.4	88.6	0.8	1.6	72.2	3.1	6.8	39.9	11.6
CMT [9]	1.4	85.2	2.8	0.9	87.6	1.8	7.2	49.8	12.6	8.1	21.8	11.8
SSE [37]	0.2	78.9	0.4	7.0	80.5	12.9	8.5	46.9	14.4	2.1	36.4	4.0
LATEM [18]	0.1	73.0	0.2	7.3	71.7	13.3	15.2	57.3	24.0	14.7	28.8	19.5
DEVISE [23]	4.9	76.9	9.2	13.4	68.7	22.4	23.8	53.0	32.8	16.9	27.4	20.9
EZSL [19]	2.4	70.1	4.6	6.6	75.6	12.1	12.6	63.8	21.0	11.0	27.9	15.8
SYNC [20]	7.4	66.3	13.3	8.9	87.3	16.2	11.5	70.9	19.8	7.9	**43.3**	13.4
SAE [21]	0.4	80.9	0.9	1.8	77.1	3.5	7.8	54.0	13.6	8.8	18.0	11.8
ALE [11]	4.6	73.7	8.7	**16.8**	76.1	**27.5**	23.7	62.8	34.4	21.8	33.1	26.3
SJE [15]	3.7	55.7	6.9	11.3	74.6	19.6	23.5	59.2	33.6	14.1	30.5	19.8
AML	**12.6**	74.5	**21.5**	11.8	**89.6**	20.8	**25.7**	66.6	**37.1**	**20.0**	38.2	**26.3**

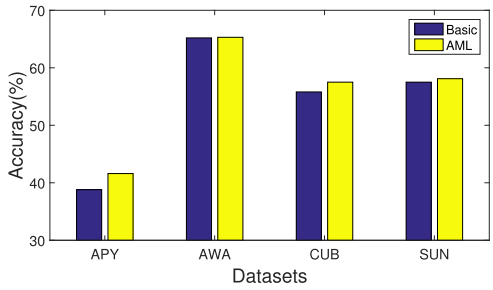

Fig. 3. Ablation study of the adaptive learning approach. 'Basic' means that no adaptive metric learning constraints are utilized ('AML' $-L_u$). 'AML' is the proposed approach.

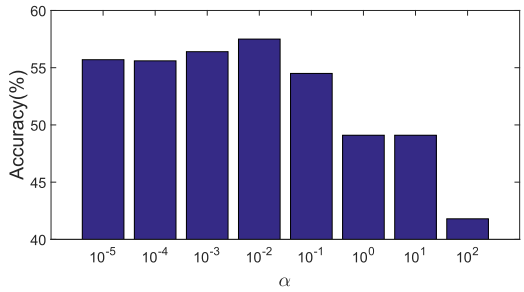

Fig. 4. The influence of adaptive learning terms on CUB under different weights.

Compared with them, our approach achieve more balanced performance between seen and unseen classes, as can be seen by the higher value of 'H'. This may attribute to the adaptive learning process, which adapt the compatibility metric to the unseen classes in the training process. Thus it helps to alleviate the overfitting problem. Compared with the linear approaches, nonlinear models are more robust in GZSL task probably because the nonlinear neural networks are more powerful to learn robust visual-semantic transformations. Moreover, [27], [28] generates image features for unseen classes and the augmented training samples make the model more robust. Therefore, applying the idea of adaptive learning to the nonlinear models may further improve the performance.

D. Effectiveness of Adaptive Learning

We exploit adaptive learning to make the model more suitable to the unseen classes. To demonstrate its effectiveness, we perform ablation studies on four datasets. The recognition results are shown in Fig. 3. We can see that the 'Basic' model has similar performance with [11], [15], since both of them utilize metric learning constraints to learn compatibility functions. After incorporating the adaptive learning constraints, the performance improves. It improves little on AWA probably due to the relative strong relations between seen and unseen classes in this dataset. The metric learned on seen classes probably have satisfied the adaptive metric learning constraints and the adaptive metric learning constraints make little influence.

E. Influence of Adaptive Learning

In order to explore how important the adaptive learning term is, we perform experiments on CUB with different values of α. The influence on the unseen classes is shown in Fig. 4. We can figure out that the model achieves the best performance when $\alpha = 0.01$ and larger weights will cause a performance drop. This indicates that the adaptive learning term plays a less important role than the original metric learning probably because the supervised information is not as accurate as that on the seen classes. However, appropriate adaption would make the model more suitable to the unseen classes.

IV. CONCLUSION

This letter proposes an adaptive metric learning approach for zero shot recognition, where a compatibility metric is learned to build up the relations between the visual space and the semantic space. In order to make the compatibility metric more suitable to the unseen classes, the class similarities are utilized to explicitly transfer the knowledge from seen-class images to similar unseen classes. Extensive experiments on four ZSL datasets show the effectiveness of the proposed approach. Our adaptive metric learning approach is based on linear embeddings. Since nonlinear models would be more powerful, we will study how to extend our approach to the nonlinear models in the future.

REFERENCES

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

[2] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[6] X. Zhu, D. Anguelov, and D. Ramanan, "Capturing long-tail distributions of object subcategories," in *Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 915–922.

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

[8] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. Comput. Vis.Pattern Recognit.*, 2009, pp. 1778–1785.

[9] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 935–943.

[10] M. Norouzi *et al.*, "Zero-shot learning by convex combination of semantic embeddings," in *Proc. Int. Conf. Learn. Represent.*, 2014.

[11] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Proc. Comput. Vis. Pattern Recognit.*, 2013, pp. 819–826.

[12] H. Jiang, R. Wang, S. Shan, Y. Yang, and X. Chen, "Learning discriminative latent attributes for zero-shot classification," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 4233–4242.

[13] J. Qin, Y. Wang, L. Liu, J. Chen, and L. Shao, "Beyond semantic attributes: Discrete latent attributes learning for zero-shot recognition," *IEEE Signal Process. Lett.*, vol. 23, no. 11, pp. 1667–1671, Nov. 2016.

[14] M. Meng and X. Zhan, "Zero-shot learning via low-rank-representation based manifold regularization," *IEEE Signal Process. Lett.*, vol. 25, no. 9, pp. 1379–1383, Jul. 2018.

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

[16] J. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov, "Predicting deep zero-shot convolutional neural networks using textual descriptions," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 4247–4255.

[17] R. Qiao, L. Liu, C. Shen, and A. V. D. Hengel, "Less is more: Zero-shot learning from online textual documents with noise suppression," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 2249–2257.

[18] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. A. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 69–77.

[19] B. R. Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2152–2161.

[20] S. Changpinyo, W. L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 5327–5336.

[21] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 4447–4456.

[22] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Classifier and exemplar synthesis for zero-shot learning," *CoRR*, abs/1812.06423, 2018.

[23] A. Frome *et al.*, "Devise: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[24] H. Jiang, R. Wang, S. Shan, and X. Chen, "Learning class prototypes via structure alignment for zero-shot recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 121–138.

[25] F. Sung, Y. Yang, X. Lin, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.

[26] S. Liu, M. Long, J. Wang, and M. I. Jordan, "Generalized zero-shot learning with deep calibration network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 2009–2019.

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

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

[29] A. Mishra, M. S. K. Reddy, A. Mittal, and H. A. Murthy, "A generative model for zero shot learning using conditional variational autoencoders," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 2269–2277.

[30] M. Bucher, S. Herbin, and F. Jurie, "Hard negative mining for metric learning based zero-shot classification," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 524–531.

[31] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2332–2345, Nov. 2015.

[32] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Comput. Neural Syst. Tech. Rep. CNS-TR-2011-001, 2011.

[33] G. Patterson, C. Xu, H. Su, and J. Hays, "The SUN attribute database: Beyond categories for deeper scene understanding," *Int. J. Comput. Vis.*, vol. 108, no. 1/2, pp. 59–81, 2014.

[34] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning—The good, the bad and the ugly," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 3077–3086.

[35] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.*, vol. 45, no. 1, pp. 503–528, Aug. 1989.

[36] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 4166–4174.

[37] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 52–68.