CORRELATED DUAL AUTOENCODER FOR ZERO-SHOT LEARNING

Ming JIANG[1], Zhiyong LIU[2]*, Pengfei LI[3], Min ZHANG[4], Jingfan TANG[5]

In the existing researches on zero-shot learning, people focus more on the mapping relationship between the visual features of images and the semantic features of each class. However, these features themselves affect the final identification for classification in a very significant way. Particularly, concerning the semantic features, some representations are relatively close to each other in some similar categories, which indicates the distinction among categories will not be so apparent. Additionally, features will also witness a redundancy if the wider span of the category appears. Therefore, to obtain more discriminative and finer-grained semantic features, this paper proposes a model on framework of the correlated dual autoencoder. Although these autoencoders are established for visual and semantic features, the two autoencoders are still related to each other without any independence. Hence, we encode the visual features, and these are added to the encoded semantic features with the decoding of added semantic features. Finally, the decoded and the original semantic features are added for better attainment and completion of semantic features. In this paper, experiments were carried out on the AwA and Cub datasets, and higher accuracy was achieved in the final classification identification.

Keywords: image classification; transfer learning; image recognition

1. Introduction

In recent years, with the development of deep learning technology, the use of computer vision for image classification and recognition has achieved good results [1, 2]. For example, in a large database such as ImageNet [3], the accuracy of recognition is already very high. However, it can be noted that the identification of these images usually requires us to collect and label a large number of images for each category. This approach severely limits the scalability of the entire model, as it becomes difficult to meet all of these needs as the size of the identification task continues to grow, and the fine-grained classification requirements increase. For example, it is easier to collect pictures of common

[1] Prof., Computer Science and Technology, HangZhou DianZi University, China
[2] * M.S., Computer Science and Technology, HangZhou DianZi University, China, corresponding author: e-mail: 604256300@qq.com
[3] Ph.D., Computer Science and Technology, HangZhou DianZi University, China
[4] Lec., Computer Science and Technology, HangZhou DianZi University, China
[5] A/Prof., Computer Science and Technology, HangZhou DianZi University, China

animals and plants such as cats and dogs, but for some rare or endangered animals and plants, it is difficult to collect a large number of these pictures in advance. This means that there may be no such data sets in the training sample. In contrast, humans are very good at identifying objects without seeing any visual images. This ability is also known as zero-shot learning. For example, a child has only seen a horse before, and then you let him go to the zoo to find a zebra. You tell him that the zebra is a horse, but there are black and white stripes on his body, so he can easily recognize the zebra. Inspired by human zero-shot learning ability, people hope that machines can also have this ability. This way of learning can extend the visual recognition of visible classes to invisible classes without additional data sets [4,5,6,7,8,9,10,11,12,13].

Zero-shot learning ultimately needs to establish a mapping relationship between visual features and semantic features, so these features themselves will have a significant impact on the final classification results. In particular, semantic features, these artificially defined attribute features or textual descriptions, have a limited distinction between similar categories, such as horses and zebras that are very close in the representation of attribute feature vectors. Besides, if the category of the category to be identified is relatively broad, such as a data set of various categories such as animals, vehicles, plants, furniture, etc., it has the problem of feature redundancy in the representation of the attribute vector because some features will only be apparent in some specific categories. Both of the above problems have caused difficulties in our following classification and identification. Therefore, this paper proposes a feature extraction model to mine these attribute features to obtain more distinguishing semantic features. Experiments show that using this method to obtain semantic features, the distinction between them is better, and higher accuracy is achieved in the following classification and recognition.

In this article we present a model architecture for a correlated dual autoencoder, where the autoencoder differs from a traditional autoencoder [14]. Its role is not only to reduce the dimension of the main components of the original information. We encode and decode visual and semantic features separately. In order to obtain more discriminative semantic features, we add the features obtained by visual feature coding to the encoded semantic features and add this new constraint to the encoded semantic features, so that it is affected by this potential visual feature. This ultimately affects the semantic feature of the decoder reconstruction. Finally, we add the decoded semantic features and the original semantic features to obtain the final semantic features. Experiments show that the semantic features obtained by this method are more distinguishable than the original semantic features.

2. Related Work

Zero-shot learning includes semantic space and visual space. Visual space is often learned through deep neural network learning and training. The most widely used semantic space is the attribute space, which requires manual attribute annotation for each class. However, for some problems between classes and classes that are relatively large or have a large number of categories, there may be some difficulties in the annotation of attribute features. Therefore, for some models, their semantic space may also contain word vectors corresponding to class names and textual descriptions of these classes [15]. Only the attribute space is used in the experiments in this article.

In order to achieve the final classification recognition effect, we need to establish a mapping relationship between semantic space and visual space. The mapping between the two can be roughly divided into these three categories. (1) Semantic space is mapped to visual space. (2) The visual space is mapped to the semantic space. (3) Projecting visual features and semantic features into the intermediate space. When choosing the mapping relationship, we need to pay attention to the hubness problem in zero-shot learning. This is actually a problem inherent in high-dimensional space: in high-dimensional space, some points become the nearest neighbors of most points. For the above mapping relationships, [16] found that the mapping from semantic space to visual space achieved the best results, because it can well alleviate the hubness problem in the high-dimensional space. Therefore, in this paper, we use the mapping from semantic space to visual space in the choice of the final mapping relationship.

The autoencoder has a good effect on data noise reduction and dimensionality reduction and can extract various good features. Because of the simplicity and effectiveness of the autoencoder, it has a wide range of applications in the field of computer vision. The autoencoder mainly includes two processes of encoding and decoding.it encodes the original features, extracts the optimized features S, and then decodes the S to obtain the reconstructed features. We hope that this reconstructed feature and the original feature will be as similar as possible. There are also examples of using autoencoders in the zero-shot learning field, such as [17]. Because the semantic space is usually lower than the dimension of the visual feature space, it uses an autoencoder for the visual features. It is hoped that the features obtained after encoding are the same as the semantic features, so as to obtain the mapping relationship between the two. In our model, we use an autoencoder for both visual and semantic features and add the encoded visual features to the encoded semantic features. The purpose here is not to construct mapping relationships but to extract better semantic features.

There is a problem with domain drift in zero-shot learning. The domain drift represents the same attribute, and its corresponding visual representation may

be different in different categories. For example, horses and pigs have the tail attribute, but they have different performance on the tail. The horse's tail is longer, and the pig's tail is shorter. If the horse is a training set and the pig is a test set, it is difficult to classify the pig using the model trained by the horse. Because the data in the training set may describe the "tail" differently from the data in the test set, although they all have the "tail" attribute, the direct migration may be biased.

The dual autoencoder model constructed by our model adds visual features to the reconstruction of semantic features, and the optimized semantic features can better represent the correct semantics of the category, to reduce the influence caused by domain drift. Previous work on zero-shot learning focused more on finding mappings, ignoring the study of attributes themselves. This paper focuses more on attributes, simply using autoencoders to optimize attribute features. In the course of the experiment, our mapping relationship uses the method mentioned in [16], and the final result is better compared with it.

3. Approach

Our model consists of three submodels. The three models are visual feature autoencoder, visual semantic correlation autoencoder, and mapping model. The coded visual features are obtained by visual feature autoencoder, which is used in visual semantic correlation autoencoders to train optimized attribute features. Optimized attribute features are trained as inputs to the mapping model, ultimately achieving the purpose of classification recognition. More details will be described below.

3.1 Problem Settings

In zero-shot learning, we are given I training set classes(denoted as Y_{tr}) and J test set classes(denoted as Y_{te}), where the training set and test set classes are disjoint, i.e. $Y_{tr} \cap Y_{te} = \emptyset$.We use the index $\{1, \dots, I\}$ to represent the training set classes and$\{I+1, \dots, I+J\}$ to represent the test set classes. The source classes contain P labeled images $\mathcal{D} = \{(v_i, y_i) | v_i \in V, y_i \in Y_{tr}\}_{i=1}^{I}$. V represents the visual feature space. Semantic information $A = \{a_c\}_{c=1}^{I+J}$ is provided for each class $c \in Y_{tr} \cup Y_{te}$. The goal of ZSL is to learn visual classifiers of test set classes $f_{zsl}: V \rightarrow Y_{te}$.

3.2 Visual feature autoencoder

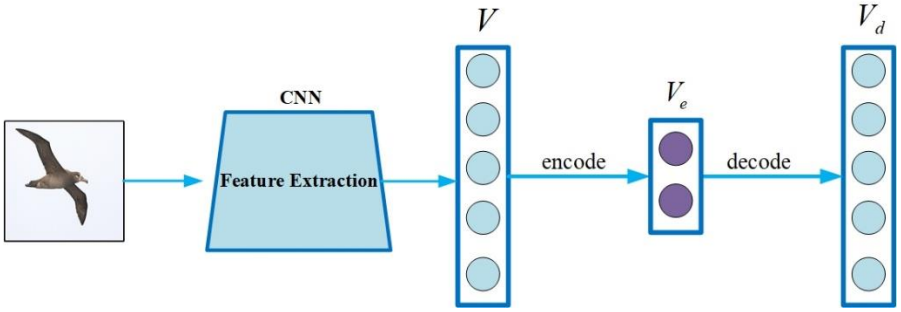

Fig. 1. The framework of visual feature autoencoder. V is the visual feature extracted from the original image using the deep neural network(VGG and Inception-V2), and we train it to construct the autoencoder. The purpose is to obtain the encoded visual feature V_e. We believe that the trained V_e is an important feature of the image. V_d represents the visual features after decoding reconstruction.

As shown in Fig. 1, we first use the deep neural network (VGG and Inception-V2) to extract features from the original image to obtain the initial visual feature vector $V \in \mathbb{R}^{B\times M}$. An autoencoder is then constructed for the initial visual feature vector V. The method of constructing the autoencoder can be completed by the following method:

$$V_e = \sigma(VW_1 + b_1) \quad (1)$$

$$V_d = \sigma(V_eW_2 + b_2) \quad (2)$$

where $V_e \in \mathbb{R}^{B\times N}$ denotes the encoded features of visual features, $V_d \in \mathbb{R}^{B\times M}$ denotes the visual features of V_e after decoding and reconstruction, $W_1 \in \mathbb{R}^{M\times N}$and $W_2 \in \mathbb{R}^{N\times M}$ are the weights which will be learned, b_1 and b_2 are bias. $\sigma(\cdot)$is the ReLU activation function.

The loss function can be formulated as:

$$L_1 = \|V - V_d\|^2 \quad (3)$$

3.3 Visual semantic correlation autoencoder

As shown in Fig. 2, we use the first visually trained visual feature to obtain a coded feature V_e from the encoder and an initial attribute feature $A \in \mathbb{R}^{B\times K}$ to construct a visual semantic correlation autoencoder model.

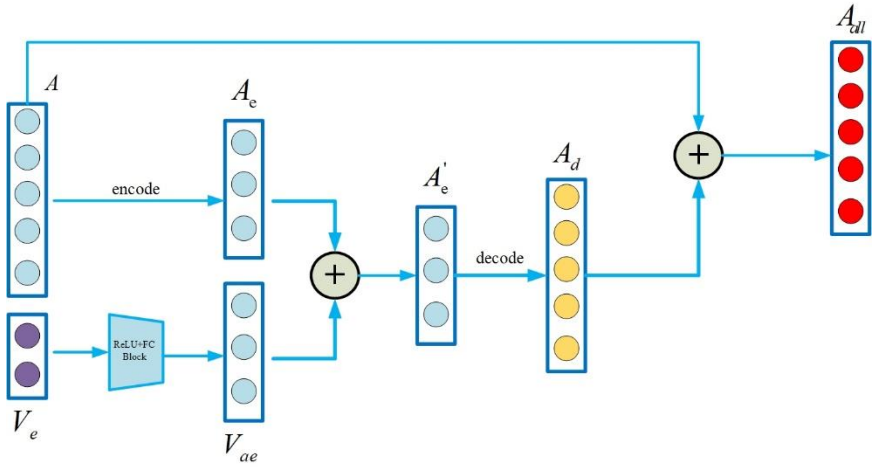

Fig. 2. The framework of visual semantic correlation autoencoder. We decode the initial attribute feature A to obtain the encoded attribute feature A_e. At the same time, V_e is mapped to the same dimensional space as A_e by ReLU+FC Block (two ReLU+FC activation function layers) to obtain the eigenvector V_{ae}. V_e represents the encoded visual features obtained from the visual feature autoencoder. Then add A_e and V_{ae} to get A'_e, and finally decode A'_e to get A_d. In order to obtain a more complete semantic expression, A and A_d are added to obtain the finally optimized attribute feature A_{all}.

We encode the initial attribute feature A to obtain the encoded attribute feature $A_e \in \mathbb{R}^{B\times L}$, and also map V_e to the space of the same dimension as A_e through two ReLU+FC activation function layers to obtain the feature vector. $V_{ae} \in \mathbb{R}^{B\times L}$. These two processes can be represented by the following formula:

$$A_e = \sigma(AW_3 + b_3) \tag{4}$$

$$V_{ae} = \sigma(\sigma(V_eW_4 + b_4)W_5 + b_5) \tag{5}$$

where $W_3 \in \mathbb{R}^{K\times L}$, $W_4 \in \mathbb{R}^{N\times C}$ and $W_5 \in \mathbb{R}^{C\times L}$ are the weights which will be learned, b_3, b_4 and b_5 are bias. $\sigma(\cdot)$ is the ReLU activation function.

We add the eigenvectors V_{ae} and A_e to get $A'_e \in \mathbb{R}^{B\times L}$, A'_e is influenced by both visual and semantic aspects, can be better expressed, and then decodes A'_e, the calculation formula is:

$$A'_e = A_e + V_{ae} \tag{6}$$

$$A_d = \sigma(A'_eW_6 + b_6) \tag{7}$$

where $A_d \in \mathbb{R}^{B\times K}$ denotes reconstructed attribute features, $W_6 \in \mathbb{R}^{L\times K}$ is the weights which will be learned, b_6 is bias. $\sigma(\cdot)$ is the ReLU activation function.

The loss function can be formulated as:

$$L_2 = \|A - A_d\|^2 \tag{8}$$

Finally, in order to express the attribute features more fully, we add the initial attribute feature A and the reconstructed attribute feature A_d to get $A_{all} \in \mathbb{R}^{B\times K}$:

$$A_{all} = A + A_d \tag{9}$$

3.4 Mapping model

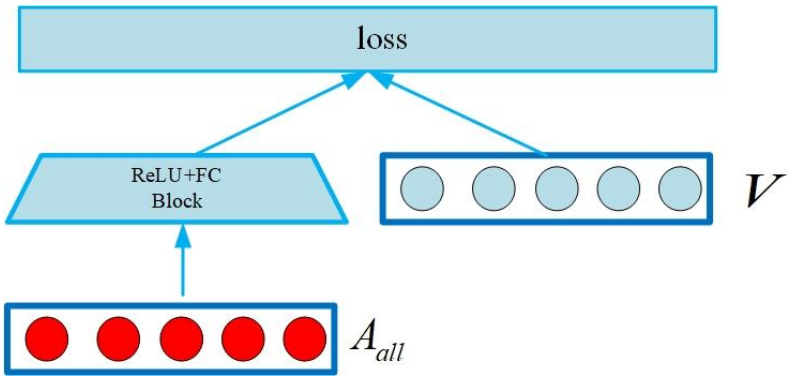

Fig. 3. The framework of mapping model. The feature vector A_{all} is mapped into the dimensional space of the feature vector V by ReLU+FC Block (two ReLU+FC activation function layers). A_{all} is an optimized attribute feature obtained by a visual semantic correlation autoencoder, and V represents an initial visual feature.

The mapping model we borrowed from the method proposed in [16]. As shown in Fig. 3, the right side is the initial visual feature V trained by the deep neural network. This D-dimensional visual feature space will be used as an embedding space that will embed the image content and semantic representation of the class to which the image belongs. In the left part, [16] uses the initial semantic feature vector A, and our model uses the optimized semantic feature vector A_{all}. A_{all} obtains a D-dimensional semantic embedding vector through two ReLU+FC activation function layers. The loss is then calculated for the visual feature vector and the semantic feature vector using the least squares method on the visual feature space. The loss function can be formulated as:

$$L_3 = \|\sigma(\sigma(A_{all}W_7 + b_7)W_8 + b_8) - V\|^2 \tag{10}$$

where $W_7 \in \mathbb{R}^{L\times M}$ and $W_8 \in \mathbb{R}^{M\times D}$ are the weights which will be learned, b_7 and b_8 are bias. $\sigma(\cdot)$is the ReLU activation function.

In the process of testing, calculate the visual distance v_i of an image and the Euclidean distance of all the optimized attribute features in the test set. The one with the smallest distance is the category we predict to belong to.

4. Experiment

4.1 Dataset and settings

Datasets: We use the classic zero-shot learning datasets of AwA and CUB in this experiment. AwA data set [18] contains 30,745 pictures of animals, with 50 categories, of which 40 are training sets and 10 are test sets. CUB data set [19] contains 11,788 images of birds, with 200 categories, 150 as a training set and 50 as a test set.

Semantic space: For AwA datasets, we use 85-dimensional continuous attribute vectors. For CUB, we use a 312-dimensional continuous attribute vector. It should be noted that only attribute features are used in this experiment, and no additional word vector features or text features are added.

Model setting and training: Experiment uses the pytorch deep learning framework to assist the training and learning model.For initial visual feature vectors, we use VGG and Inception-V2 networks to train and extract them. The extracted visual feature vector size is 1024 dimension. Adam is used to optimise our model with a learning rate of 0.0001.

Parameter setting: Weight matrix of two FC layers in the construction of visual feature autoencoder, $W_1 \in 1024 \times 700$， $W_1 \in 700 \times 1024$。The weight matrix involved in the construction of the visual semantic correlation autoencoder, in the data set CUB, $W_3 \in 312 \times 200$， $W_4 \in 700 \times 500$, $W_5 \in 500 \times 200$, $W_6 \in 200 \times 312$, in the data set AwA, $W_3 \in 85 \times 50$, $W_4 \in 700 \times 500$， $W_5 \in 500 \times 50$, $W_6 \in 50 \times 85$. The weight matrix of two FC layers in the mapping model, in the data set CUB, $W_7 \in 312 \times 700$, $W_8 \in 700 \times 1024$， in the data set AwA, $W_7 \in 85 \times 700$， $W_8 \in 700 \times 1024$。

4.2 Experimental results on the AwA dataset and CUB dataset

Since these two data sets are relatively classical in the field of zero-shot learning and are relatively small in scale, a lot of work has been done to deal with them and some achievements have been made. As shown in Table 1, we have selected some representative results for comparison.

Comparative results on AwA: Using our model, 86.935% accuracy can be achieved on AwA dataset. This improves the accuracy by 0.2% compared with DEM method which directly uses original semantic space for mapping.

Comparative results on CUB: On the CUB dataset, our model can be 1.1% more accurate than the DEM model, reaching 59.4181%.

Table 1

Compare accuracy with CUB on the dataset AwA. SS in the table represents the semantic space, where A represents the attribute space, W represents the word vector space, and D represents the text description space (valid only for CUB data set). F represents the method of obtaining visual space, where F_O represents the overfeat method; F_G represents the GoogLeNet network structure; and F_V represents the VGG network structure. N_G represents the Inception-V2.

Model	F	SS	AwA	CUB
AMP[11]	F_O	A+W	66.0	-
SJE[7]	F_G	A	66.7	50.1
SJE[7]	F_G	A+W	73.9	51.7
ESZSL[6]	F_G	A	76.3	47.2
SSE-Relu[4]	F_V	A	76.3	30.4
JLSE[20]	F_V	A	80.5	42.1
SS-Voc[10]	F_O	A/W	78.3/68.9	-
SynC-struct[21]	F_G	A	72.9	54.5
SEX-ML[13]	F_V	A	77.3	43.3
DeViSE[5]	N_G	A/W	56.7/50.4	33.5
Socher et al.[22]	N_G	A/W	60.8/50.3	39.6
MTMDL[23]	N_G	A/W	63.7/55.3	32.3
Ba et al.[24]	N_G	A/W	69.3/58.7	34.0
DS-SJE[15]	N_G	A/D	-	50.4/56.8
DEM[16]	N_G	A/W(D)	86.7/78.8	58.3/53.5
Ours	N_G	A	86.9	59.4

4.3 Experimental results using different feature maps

As shown in table 2, we mapped different attribute features and initial visual features. The experimental results are as follows:

$\mathbf{A} \rightarrow \boldsymbol{V}$: Map the initial attribute features to the initial visual features. This is what [16] does. The accuracy rate of 58.3% was obtained in CUB data set and 86.7% accuracy in the AwA dataset.

$\boldsymbol{A_d} \rightarrow \boldsymbol{V}$: The reconstructed attribute feature is mapped to the initial visual feature. Through experiments, we found that this kind of mapping method has different effects on different data sets. For CUB datasets, the effect is not as good as $\mathbf{A} \rightarrow \mathbf{V}$, with 57.5% accuracy. For AwA datasets, the effect is similar to $\mathbf{A} \rightarrow \mathbf{V}$, with 86.6% accuracy.

$\boldsymbol{A_{all}} \rightarrow \boldsymbol{V}$: The final experiment proves that this mapping method of remapping the reconstructed attribute features plus the initial attribute features to the initial visual features is the best.It has had good results in both the AwA and

CUB data sets. The accuracy rate of 59.4% was obtained in CUB data set and 86.9% accuracy in the AwA dataset.

The experimental results show that the attribute features obtained after the reconstruction are added to the original attribute features have better semantic representation. The added attribute features not only have the original attribute features but also contain additional hidden information of visual feature blessing, which makes the expression of attribute features more complete and more granular.

Table 2

Results of different feature mappings.

	CUB	AwA
$A \rightarrow V$	58.3	86.7
$A_d \rightarrow V$	57.5	86.6
$A_{all} \rightarrow V$	59.4	86.9

5. Conclusion

In this paper, we propose a novel autoencoder model for zero-shot learning. This model mainly contains three submodels. The visual feature autoencoder model is to obtain the encoded visual features. The visual semantic correlation autoencoder model plays the role of optimizing semantic features. It combines the coded visual features generated by the previous model so that the semantic features are affected by the visual features so that the semantic features can be described better and more completely. The last model is the mapping model. We map the optimized semantic features to the visual feature space and then classify and identify them according to the Euclidean distance. Our model optimizes semantic features to improve the accuracy of the final classification. Extensive experiments on two benchmark datasets show the superiority of the proposed approach.

Acknowledgements

This work is supported by Zhejiang Provincial Technical Plan Project (No. 2019C03096, 2018C03039).

REFERENCES

[1] *A. Krizhevsky, I. Sutskever, and G. E. Hinton*, Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, **vol. 1**, Dec. 2012, pp. 1097-1105

[2] *P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun*, Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229, 2013

[3] *O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei*, ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), **vol. 115**, no. 3, Sept. 2015, pp. 211-252

[4] *Z. Zhang and V. Saligrama*, Zero-shot learning via semantic similarity embedding, In Proceedings of the IEEE International Conference on Computer Vision, Dec. 2015, pp. 4166-4174

[5] *A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al.*, Devise: A deep visual-semantic embedding model. In Advances in neural information processing systems, **vol. 2**, Dec. 2013, pp. 2121-2129

[6] *B. Romera-Paredes and P. H. Torr*, An embarrassingly simple approach to zero-shot learning. Proceedings of the 32nd International Conference on Machine Learning (ICML), **vol. 37**, Jul. 2015, pp. 2151-2161

[7] *Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele*, Evaluation of output embeddings for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2015, pp. 2927-2936

[8] *M. Thomas, V. Jakob, P. Florent, and C. Gabriela*, Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In Computer Vision-ECCV 2012, Oct. 2012, pp. 488-501

[9] *Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid*, Labelembedding for attribute-based classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2013, pp. 819-826

[10] *Y. Fu and L. Sigal*, Semi-supervised vocabulary-informed learning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016

[11] *Z. Fu, T. Xiang, E. Kodirov, and S. Gong*, Zero-shot object recognition by semantic manifold distance. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2015, pp. 2635-2644

[12] *M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean*, Zero-shot learning by convex combination of semantic embeddings. In ICLR, Apr. 2014, pp. 488-501

[13] *M. Bucher, S. Herbin, and F. Jurie*, Improving semantic embedding consistency by metric learning for zero-shot classiffication. In European Conference on Computer Vision, Oct. 2016, pp. 730-746

[14] *D. E. Rumelhart, G. E. Hinton, and R. J. Williams*, Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Learning Internal Representations by Error Propagation. MIT Press, Cambridge, MA, USA, 1986, pp. 318-362

[15] *S. Reed, Z. Akata, B. Schiele, and H. Lee*, Learning deep representations of fine-grained visual descriptions. In CVPR, Jun. 2016

[16] *L. Zhang, T. Xiang and S. Gong*, Learning a Deep Embedding Model for Zero-Shot Learning. In CVPR, Jul. 2017

[17] *E. Kodirov, T. Xiang and S. Gong*, Semantic Autoencoder for Zero-Shot Learning. In CVPR, Jul. 2017

[18] *C. H. Lampert, H. Nickisch, and S. Harmeling*, Attributebased classification for zero-shot visual object categorization. PAMI, **vol. 36**, no. 3, Mar. 2014, pp. 453-465

[19] *C. Wah, S. Branson, P. Perona, and S. Belongie*, Multiclass recognition and part localization with humans in the loop. In ICCV, Nov. 2011

[20] *Z. Zhang and V. Saligrama*, Zero-shot learning via joint latent similarity embedding. In CVPR, Jun. 2016, pp. 6034-6042

[21] *S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha*, Synthesized classifiers for zero-shot learning. In CVPR, Jun. 2016
[22] *R. Socher, M. Ganjoo, C. D. Manning, and A. Ng*, Zero-shot learning through cross-modal transfer. In NIPS, **vol. 1**, Dec. 2013, pp. 935-943
[23] *Y. Yang and T. M. Hospedales*, A unified perspective on multi-domain and multi-task learning. In ICLR, May. 2015
[24] *J. Lei Ba, K. Swersky, S. Fidler, and R. Salakhutdinov*, Predicting deep zero-shot convolutional neural networks using textual descriptions. In ICCV, Dec. 2015, pp. 4247-4255