Journal of Science (JOS) ISSN 2324-9854
Vol. 5, No. 1, 2016, Pages: 455-460
Copyright World Science Publisher, United States
www.worldsciencepublisher.org

A survey on Zero-shot learning

Fengyi Song[a], Yi Chen[a,b]

[a]School of Computer Science and Technology, Nanjing Normal University, Nanjing, Jiangsu 210023, China
[b]KeyLaboratoryofImageandVideoUnderstandingforSocialSafety(NanjingUniversityofScienceandTechnology),Nanjing,210094,China

Abstract

Zero-shot learning (ZSL) has recently received extensive attention for its potential in achieving scalable object recognition with lower human labor cost relative to traditional supervised learning. However, zero-shot learning is a nontrivial problem, and its feasibility relies on satisfaction of several important assumptions and conditions, where learning knowledge shareable between seen classes and unseen classes becomes the foundation. A plenty of works are proposed from different views with various formulations while obeying the foundation. We will review the literature of zero-shot learning comprehensively while putting emphasis on analyzing their motivations, assumptions, and exact mechanism for learning transferable knowledge that is helpful for connecting testing images and description of unseen classes. Finally, benchmarks for evaluating and comparing kinds of approaches are discussed mainly involving the datasets, protocols and evaluation measures. We hope this review may shed light on advanced solutions to zero-shot learning.

Keywords: Zero-shot learning, Knowledge transfer, Survey

1. Introduction

Zero-shot learning (ZSL) refers to a special object recognition paradigm where partial classes do not have any training image available but with some semantic descriptions of these classes, and then at the test phase try to recognize the images of these never seen classes. Besides, there usually provide some auxiliary classes with a number of labeled images and also their semantic descriptions. The success of zero-shot learning will promote large-scale object recognition for modern machine vision systems, which is very important for real-scenario applications such as automatic driving system, where there are tremendous unexpected objects and scenes need to be recognized for benefitting the right decision control.

Despite great success have been made on general supervised learning paradigm with longtime improving techniques as feature representation learning, large capacity and flexible model training, and also availability of large image data and economical computation resource, some inherent limitations of supervised learning make it hard for implementing large-scale object recognition system, such as the involved great manual labeling cost. Alternatively, zero-shot learning will greatly improve modern machine vision system with better autonomic and adaptability.

However, zero-shot learning is a nontrivial problem. There is some fundamentals assumptions and motivation for zero-shot learning to work reasonably. Generally, three conditions should be satisfied for a feasible zero-shot recognition system, i.e., (1) vector representation of objects classes should be provided either by human labeling or automatic mechanism, which is also called label embedding; (2) the embedding representation is competent for transferring knowledge from seen classes to unseen classes; (3) these embedding representation should be closely related to visual features in image domain, which ensures the convenience for implementing the mapping from image domain to the embedding space, i.e., image embedding. We can see that the embedding space bridges the semantic gap between low-level image features and high-level object classes.

Attributes as an intermediate representation satisfy these conditions well and become a popular knowledge transferring media suitable for zero-shot learning [1, 2, 3]. For example, we do not have any images of the "aye-aye" for training but we know how it looks alike –"it is a nocturnal animal living in trees, having large eyes and having long middle fingers". Given a testing image, we can justify whether it contains an "aye-aye" by comparing the visual appearance with these descriptions (visual attributes). Following this approach, many works focus on several key issues including enhancing the generalizable ability of attribute extractors[4], or exploiting the end-to-end learning strategy with compatible function formalization [5, 6] that estimating the consistency between the label embedding and images, or addressing project domain shifting problem in attribute learning [7].

Beyond semantic attributes representation for object classes embedding, other representations with less human labor cost are also explored, such as word vector representation that can be automatically obtained by exploiting unsupervised text mining technique on massive unstructured text corpus [8, 9]. Such representation can also relate unseen object classes with seen object classes by leveraging statistic information about the neighboring texts related to classes' names. Recently many works appeared using such representation since their low cost, flexibility in large-scale recognition and also model design (not need to directly learn the mapping from images to embedding space).

Besides these two approaches for class description, directly

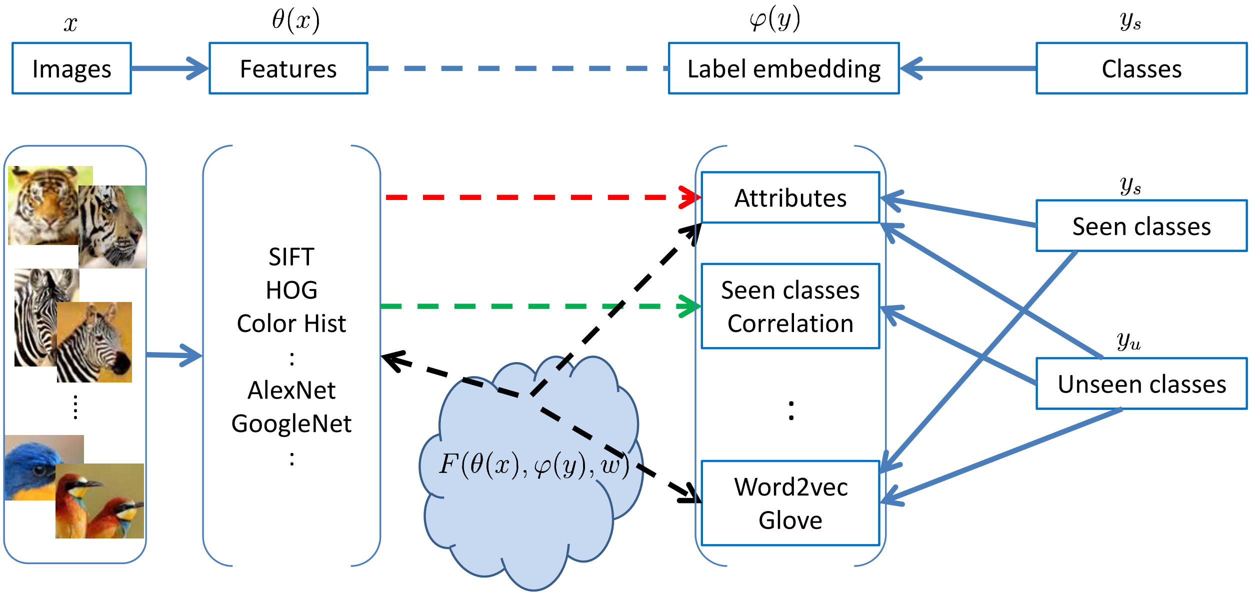

Figure 1: Summarization of typical zero-shot learning paradigms, where the solid lines represent the known relations and the dash lines represent the relationships need to be learned and the arrow direction shows their dependence relation. Typical paradigms of ZSL are distinguished in the implementation for bridging images with classes descriptions, where the red line indicates the two-order transformation approach, the green line related to the direct seen classes correlation approach, and the dark lines indicate one-order and higher-order transformation approach (best viewed in the electronic form).

relate the unseen object classes to the seen object classes also a reasonable way for zero-shot learning, since object classes hierarchical relationships are commonly exist and can be obtained through knowledge database at hand, including lexical dictionary as WordNet [10]. For example, Hoo and Chan [11] utilized hierarchical class concept for relating the unseen classes to the seen classes. Zhang and Saligrama [12] proposed to project each unseen object and testing instance to the subject spaces of the seen classes, i.e., either unseen classes or unseen testing instance are expressed as a histogram of seen class proportions.

By far, we have introduced the scenario of zero-shot learning, summarized the assumption and motivations for designing a feasible zero-shot learning systems, and emphasized the key issue is to learn transferable knowledge for bridging the semantic gap between images and objects classes. As we have investigated several representations media for knowledge transfer in zero-shot learning, next, we will make a comprehensive survey of the literature working on these media with the emphasis on analyzing their motivation and approaches achieving better knowledge transferring from the seen classes to the unseen classes.

2. Typical approaches to zero-shot learning

A general pipeline of zero-shot learning can be illustrated in Fig.1 We can see that the key issue for zero-shot learning lies on the bridge connecting the huge semantic gap between images and object classes. There are various detailed formulation of the bridge, and can be generally summarized as the following three categories. We will first introduce the commonly adopted attribute-based zero-shot learning approach in data transformation view, i.e. what we called *Two-order Transformation*. Next, we will discuss two adverse evolution of this approach also from the view of data transformation, with one direction to simplify data transformation by directly justifying whether image and object label embedding vectors are matched or not, and is called *One-order transformation*; the other is to make more complicated data transformation for disentangling sharable factors better explaining the inherent relationship between images and object labels, and is called *High-order transformation*.

2.1. Two-order transformation approach

By projecting both the images and objects labels into a shared common space such as commonly explored semantic attribute space, zero-shot object recognition can be implemented by simply comparing their similarity between their new representation. This is the typical pipeline in what we called *Two-order transformation*. We can see that attribute as an intermediate representation bridges the great semantic gap between images and object classes. For an unseen class, given its attribute description, then we can justify the class of a testing image by computing its attributes response. In some extent, zero-shot learning degrades to the attribute prediction task.

Typical methods include direct attribute prediction (DAP) and indirect attribute prediction (IAP) [13]. In DAP, attribute predictors are learned on labeled training data of the seen object classes. Then expecting these attribute predictors can work well on images of any classes including both the seen classes and the unseen classes. Finally, Lampert et al. [13, 3] show a Bayes probabilistic inference framework for predicting any objects classes by incorporating both the predicted attributes and pre-specified attribute representation for each class. Alternatively, in IAP framework, they do not try to learn attribute predictors directly, but to learn seen classes classifiers by traditional supervised learning techniques since labeled images of the seen classes are provided. Then, applying these learned seen object classifier on any testing images and obtaining the posterior distribution of the seen classes given the testing image. Next, using these posteriors as weights for accumulating all the attributes induced from the class-attribute matrix. Similar to IAP, Mohammad Norouzi et al. [14] use posterior probabilities of the seen classes given a testing image as the weights to convexly combine the label embedding of the seen classes, and finally formed the image embedding in the same space for label embedding, then zero-shot learning can be achieved by simply comparing the embedding representation in the same space for images and classes. The reasonability and success of such methods may heavily rely on the manifold assumption of the embedding space, i.e., similar classes should have similar embedding representation, which seems hold true for the word2vec embedding space [15]. However, such an assumption may not always be well satisfied for any embedding spaces. With this insight in mind, while following the similar way as [14] but going step further, Changpinyo et al. [16] introduce a latent dictionary base (they called “phantom” object classes) in both embedding space and the model parameter space, and consider to align the embedding space with model space by constraining consistent neighbor relationship in both these two spaces.

2.1.1. Special issues in two-order transformation approach

In the typical paradigm of zero-shot learning introduced above including both DAP and IAP [13], we can see that learning attribute predictor and predicting final object class are two independent processes, which may not ensure optimized zero-shot learning accuracy since there usually exists great distribution difference for seen objects and unseen objects. We will investigate two important issues for addressing these deficiencies.

Jointly learning attributes with discriminant information. As we can see that DAP is actually neglect the relationships between attributes and also classes, it is necessary to explore these relationship to improve DAP approach for zero-shot learning. Wang and Ji [17] utilized Bayes Network to explore relationship between attributes and also classes for discriminant attribute learning that may benefit latter class recognition task. Mahajan et al. [18] present a new approach to learning attribute-based descriptions of objects. They do not assume that the descriptions are hand-labeled, and on the contrary, they jointly learn both the attribute classifiers and the descriptions in the data-driven fashion. By incorporating class information into the attribute classifier learning, they get an attribute level representation that generalizes well to both unseen examples of known classes and also unseen classes. Similarly, Liang et al. [19] build a multiplicative framework for attribute learning in a shared space of both object class and images.

How to obtain the vector embedding of classes? In general zero-shot learning paradigm, vector representation of object classes is supposed to be given previously. For example, popular used visual semantic attribute representation for object classes is provided manually according to experts’ knowledge. We can see that such a process is too labor costing and may introduces human bias. Consequently, researchers consider learning vector representation of categories automatically. In zero-shot learning scenario, since we already know the vector representation of seen objects, it is feasible to find the relationship between an unseen class and the seen classes, and inference the vector representation of this unseen class by leveraging that of seen classes [20]. Following this way, lexical hierarchies provided in WordNet [10] can also be exploited to choose closely related class for visual attributes transferring [21, 22, 23], such as ancestor classes. Alternatively, objects and attributes usually have contextual text, semantic relatedness can be measured by analyzing their contextual text correlation or co-occurrence statistic.

As Al-Halah et al. [24] point out that class-attribute relations are complex and it is hard to model them by simple co-occurrence statistic as introduced above. Since object class and attributes are essentially text word, techniques for text processing provide us effective tools for analyzing class-attribute relationship. For example, Mikolov et al. [15] provides an efficient approach to learn a vector representation for each word through mining contextual words distribution on a large text corpus such as Wikipedia, and the finally learned vector representation for words with similar context are also close in the vector space. Such techniques are exploited in [25] for building word representation for object classes, and archived impressive results on large-scale zero-shot learning. However, as Al-Halah et al. [24] pointed out that the performance of such vector representation for object classes still inferior to that of attribute representation, and the key problem may lies in the insufficient modeling of the complex relationship between classes and attributes. Finally, they proposed to model the relationship between classes and attributes directly with a bilinear model in the word embedding space [15]. The specified relationship are learned in data-driven mode with training data across different class and attributes, which ensures better inference of class-attribute representation for unseen classes in zero-shot learning.

2.2. One-order transformation approach

Unlike the previous paradigm, zero-shot learning can also be implemented by directly estimating the direct matching score between images and classes (with label embedding vector), this is the typical idea in what we called “One-order transformation”. Going further along this way, the key issue lies on implementing an effective computation of the matching score. Considering the image domain and label embedding space are commonly from two different space, it is not suitable to utilize usual similarity or distance function. Consequently, bilinear function is suitable for such case, formulated as $\phi(x)^T W\varphi(y)$, where $\phi(x)$ is feature representation of an image, and $\varphi(y)$ is label embedding vector for the class y, and W is the parameter to be learned. Several works has utilized this formalization [25, 22, 26, 6]. These works are different with respect to the specific loss criteria used for driving the best parameter of W. For example, SJE [22] used multi-class loss with the form as $\max_{y\in\mathcal{Y}}\{0, \Delta(y_n, y) + \phi(x_n)^T W\varphi(y) - \phi(x_n)^T W\varphi(y_n)\}$, and ALE [6] used a weighted approximate ranking loss with the form as $\sum_{y\in\mathcal{Y}} \max\{0, \Delta(y_n, y) + \phi(x_n)^T W\varphi(y) - \phi(x_n)^T W\varphi(y_n)\}$. DeViSE [25] used the similar loss as ALE [6] with the form of $\sum_{y\neq y_n} \max\{0, margin + \phi(x_n)^T W\varphi(y) - \phi(x_n)^T W\varphi(y_n)\}$, but optimized in the whole framework of deep learning for driving also better feature representation of images. Romera-Paredes [26] used square loss in multi-task learning for computation efficiency.

Let's revisit the formulation of $\phi(x)^T W\varphi(y)$, actually, it can be seen as a linear model on W given $\phi(x)$ and $\varphi(y)$, and can be reformulated as $Vec(W)^T Vec(\phi(x) \otimes \varphi(y))$, where $Vec(:)$ means to reshape a matrix into a vector by concatenating each column of the matrix. Naturally, there is a question arising "is a linear function enough for modeling the complicated relationship between images $\phi(x)$ and label representation $\varphi(y)$?", especially under complicated environmental variations such as pose, lighting. With this in mind, Xian et al. [27] proposed a nonlinear compatibility framework that learns a collections of linear models making the overall function piecewise linear, as follows, $F(x, y) = \max_{1\leq i\leq K} w_i^T(\phi(x) \otimes \varphi(y)) = \max_{1\leq i\leq K} \phi(x) W_i^T \varphi(y))$, where K represent the number of latent variable W_i that depicting several typical consistent visual distribution corresponding to typical pose and appearance. Further, they used the same ranking loss function as ALE [6] for driving a collection of optimized W_i. Akata et al. [28] put emphasis on specific information correspondence between part visual features and part semantic descriptions, where the parts location are exploited for extracting localized visual features. Finally, a set of compatible functions are learned for each combination between a set of visual parts and text parts , while used the same ranking loss as introduced above.

2.3. *High-order transformation approach*

While considering that complicated relationship exists between images and classes, it is necessary to exploit complex model for capturing the stable and transferable knowledge that may generalize well to the unseen classes. Recently, latent factor analysis approaches are explored in data transformation process for tackling zero-shot learning problem. Zhang and Saligrama [12] proposed a fourth order transform between objects class and image instances, i.e., source domain data, the associated further embedding of source domain data and also target domain data, the paired space of these two embedding labeled with matched or not. And they built a latent probabilistic model with Markov chain assumption for analyzing the complicated relationships between these transformed spaces. Such a complex model is expected to disentangle class-independent factors that may explain unseen classes well.

While disentangling class-independent factors, it is necessary to filter out noise contain in representations of both images and classes, especially for unsupervised word vector representation. Qiao et al. [29] consider to suppress the noise in text embedding ($\varphi(y)$) of the object class, they proposed a feature selection strategy on $\varphi(y)$ by applying a linear transform term W_z explicitly and put $L_{2,1}$ constrains on W_z for achieving noise suppression. Finally, the feature selection term for noise suppress is learned together with compatible function parameter W_x under the similar framework as [26] with the same loss as follows $\|XW_x^T W_z Z - Y\|_F$, where X is the image matrix, Z is the class label embedding, and Y is the label indicator matrix of images on seen classes.

3. Datasets and performance comparison

Since zero-shot learning is a typical object recognition paradigm, general object recognition dataset can be used for validating the effectiveness of algorithms addressing zero-shot learning. With another consideration that knowledge can be learned and transferred well between related object classes, it is suitable to carry out zero-shot learning for recognizing objects belong to similar coarse-level classes, such as animal dataset, find-grained bird species categorization dataset, cars dataset, and large-scale ImageNet dataset. We will introduce these dataset with the emphasis on summarizing their special characteristics in classes relatedness, the number of classes, and standard protocols for zero-shot learning (see Tab.1). Finally, we will investigate the performance reported on these databases.

Table 1: Statistics of the popular datasets used for validating approaches of ZSL.

Datasets	# Seen cls.	# Unseen cls.	Attributes
AwA [13, 3]	40	10	85
CUB-200-2011 [30]	150	50	312
aP&Y [31]	20	12	64
SUN [32]	645/707	72/10	102
ImageNet [33, 34]	1000	20,842	-

3.1. *Datasets*

Animals with Attributes (AwA) [13, 3]. AwA is a general animal concept dataset with most animal categories are mammals. It covers 50 animal classes with 85 handcrafted attributes annotated on class-level and consists of 30,475 images, where the minimum number of images for any class is 92 (mole) and the maximum is 1,168 (collie). Standard classes splitting for zero-shot learning is provided with fixed 40 classes as seen data and the left 10 classes as unseen data.

Caltech-UCSD CUB-200-2011 Birds (CUB-200-2011) [30]. CUB-200-2011 is a typical dataset for fine-grained recognition, i.e., to recognize detailed species of birds. It covers 200 categories of birds with 312 attributes annotated on image-level, and consists of 11,788 images with the nearly equal number(60) of images for each category. There is no standard protocol of classes splitting for zero-shot learning, and usually, researchers use 150 classes as seen data and the left 50 classes as unseen data, and the exact splitting can be done in a random way and report performance in cross-validation way.

a-Pascal & a-Yahoo (aP&Y) [31]. aP&Y covers 32 classes with large concept divergence. These objects classes including "animals", "vehicles", and "things". Each class has large distinct in concept and consists of the different number of images ranging from 150 to 1000, and along with over 5000 instances of people. Totally, 64 handcrafted attributes are explored for depicting each class annotated on image-level. There is no standard classes splitting protocol for zero-shot learning, but

Table 2: Performance comparison of typical approaches with their mean multi-classification accuracy, while numbers in bold font are the published results and numbers in normal font are the re-produced results by [16] and numbers in italic font are the re-produced results by [35] with vgg-verydeep 19 [36]. In the last column, we list F@K top-1 performance (equals to mean multi-classification accuracy) obtained under two settings, i.e., does not taking the seen classes as candidates classes and taking both the seen classes and unseen classes as class candidates for prediction, while the latter is shown in parentheses.

Methods	**Shallow feature**[3, 31, 32] **/ Deep feature** {GoogLeNet[37](AlexNet[38], vgg19[36])}				
	AwA[13]	**CUB** [30]	**aP&Y** [31]	**SUN** [32]	**ImageNet** [33, 34]
Two-order Trans.					
DAP [3]	**41.4**/60.5(50.0)	28.3/39.1(34.8)	**16.9**/-	**18.0**/44.5	-
IAP [3]	**42.2**/57.2(53.2,*57.23*)	24.4/36.7(32.7)	**19.1**/-(-, *38.16*)	**22.2**/40.8(-, *72.00*)	-
BN [17]	**43.4**/-	-	-	-	-
One-order Trans.					
ALE [5]	**37.4**/-	**18.0**/-	-	-	-
ALE [6]	**48.5**/-	**26.9**/-			
ALE [16]	34.8/53.8(48.8)	27.8/40.8(35.3)	-	-/53.8	-
SJE [22]	**42.3/66.7(61.9)**	**19.0/50.1(40.3)**	-	-	-
SJE [16]	36.2/66.3(63.3)	34.6/46.5(42.8)	-	-/56.1	-
ESZSL [26]	**49.3**/59.6(53.2,*75.3*)	37.0/44.0(37.2)	**27.27**/-(-,*24.2*)	**65.75**/-(-,*82.1*)	
Devise [25]	-	-	-	-	**0.8(0.3)**
Hight-order Trans.					
ConSE [14]	36.5/63.3(56.5)	23.7/36.2(32.6)	-	-/51.9	**1.4(0.2)**
Phantom [16]	**41.5/72.9(62.8)**	**36.4/54.7(47.1)**	-	-/62.7	**1.5(-)**
SSE-ReLU [35]	-/-(-,**76.33**)	-/-(-,**30.41**)	-/-(-,**46.23**)	-/-(-,**82.50**)	-
JLSE [12]	-/-(-,**79.12**)	-/-(-,**41.78**)	-/-(-,**50.35**)	-/-(-,**83.83**)	-

a-Yahoo containing 12 classes was initially collect for validating the generality of attribute predictor learned on a-Pascal. So it is reasonable to use the 20 classes in a-Pascal as seen data, and use the 12 classes in a-Yahoo as unseen data.

SUN Attribute (SUN) [32]. SUN is initially collected for high-level scene understanding and fine-grained scene recognition. It explored 102 discriminative attributes for depicting more than 700 categories with 14,340 images. Like CUB, there is no standard seen/unseen classes splitting for zero-shot learning, and random splitting is utilized in several works with different ratio including 645/72 used in [16] and 707/10 used in [26, 12], and finally reporting average performance on several tries.

ImageNet [33, 34]. ImageNet is a large-scale object recognition dataset and has several released versions. For example, ImageNet ILSVRC 2012 1K contains 1000 object categories, and ImageNet 2011 21K contains 20,842 object categories. The former is commonly used as the seen classes and the latter is used as the unseen classes [25, 14, 16]. This dataset is used for validating large-scale zero-shot learning.

3.2. *Performance comparison*

As different approaches are proposed for zero-shot learning, it is hard to directly compare their performance in an absolutely fair condition due to difference in representation of images and classes, and also the seen/unseen classes splitting etc. We will lists their reported performance with introduction of detailed experimental settings (see Tab.2). We can see that large-scale zero-shot learning on ImageNet has very low mean multi-classification accuracy due to the large size of classes and also its inherent difficulty, while there are other measures [25, 14, 16] for reflecting more detailed results such as Flat hit@K (F@K) and Hierarchical precision@K (HP@K) used in , and also consider to divide the testing unseen classes according to their distance to the seen classes in the label structure that is indicated by hops in WordNet synset hierarchy. In addition, when comparing performance of different approaches it is necessary to clearly introduce whether the seen classes is chosen as the candidate classes to be predicted, this is very important in large-scale zero-shot learning. [25, 14, 16] reported their performance on ImageNet as 0.3/0.8, 0.2/1.4, and -/1.5 respectively, where the formers are obtained by taking seen classes as candidate classes for predicting and it is lower than that only taking unseen classes as candidate predict classes.

4. Conclusion

Zero-shot learning trying to recognize the images of the unseen classes while given the description of the unseen classes, has been proved to be a feasible learning paradigm, since we can learn some transferable knowledge from the seen classes that is helpful for connecting the relationship between testing images and unseen classes. Consequently, there are two key issues for zero-shot learning, i.e., get the description of unseen classes and learn knowledge from seen classes for bridging the images with that description. We investigated several methods with the emphasis on analyzing motivations and solutions to these two issues, including popular attribute-based knowledge transferring mechanism as DAP and IAP, end-to-end learned compatible function of images and label embedding, and other latent analysis perspectives. We summarized these approaches from the general view of information transformation, and categorize them into three typical types, i.e., two-order transformation approach, one-order transformation approach, and high-order transformation approach. Finally, we summarize the typical datasets used for carrying out experiments for validating algorithms of zero-shot learning, and give a comparison of the reported performance on these datasets. To be noticed that, zero-shot learning is nontrivial task especially on large-scale classes, which is the potential research direction for future work.

5. Acknowledgements

This paper was supported by NSFC 61503188, 61562041, Natural Science Foundation of Jiangsu Province BK20161020, BK20150982, Program of Natural Science Research of Jiangsu Higher Education Institutions 15KJB520023,14KJB520021, KJR 16088,1602, and the Jiangsu Key Laboratory of Image and Video Understanding for Social Safety (Nanjing University of Science and Technology), Grant No. 30916014107. .

References

[1] M. Palatucci, D. Pomerleau, G. E. Hinton, T. M. Mitchell, Zero-shot learning with semantic output codes, in: Advances in neural information processing systems, 2009, pp. 1410–1418.

[2] N. Kumar, A. Berg, P. N. Belhumeur, S. Nayar, Describable visual attributes for face verification and image search, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (2011) 1962–1977.

[3] C. H. Lampert, H. Nickisch, S. Harmeling, Attribute-based classification for zero-shot visual object categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2014) 453–465.

[4] Z. Al-Halah, M. Tapaswi, R. Stiefelhagen, Recovering the missing link: predicting class-attribute associations for unsupervised zero-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[5] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for attribute-based classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 819–826.

[6] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for image classification, IEEE transactions on pattern analysis and machine intelligence 38 (2016) 1425–1438.

[7] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, Transductive multi-view zero-shot learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (2015) 2332–2345.

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, volume 26, 2013, pp. 3111–3119.

[9] J. Pennington, R. Socher, C. Manning, Glove: global vectors for word representation, in: Conference on Empirical Methods in Natural Language Processing, 2014.

[10] G. A. Miller, Wordnet: a lexical database for english, Communications of the ACM 38 (1995) 39–41.

[11] W. L. Hoo, C. S. Chan, Zero-shot object recognition system based on topic model, IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans 45 (2014) 518–525.

[12] Z. Zhang, V. Saligrama, Zero-shot learning via joint latent similarity embedding, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[13] C. H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[14] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, J. Dean, Zero-shot learning by convex combination of semantic embeddings, in: International Conference on Learning Representations (ICLR), 2014.

[15] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[16] S. Changpinyo, W. L. Chao, B. Gong, F. Sha, Synthesized classifiers for zero-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[17] X. Wang, Q. Ji, A unified probabilistic approach modeling relationships between attributes and objects, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2120–2127.

[18] D. Mahajan, S. Sellamanickam, V. Nair, A joint learning framework for attribute models and object descriptions, in: IEEE International Conference on Computer Vision, 2011, pp. 1227–1234.

[19] K. Liang, H. Chang, S. Shan, X. Chen, A unified multiplicative framework for attribute learning, in: IEEE International Conference on Computer Vision, 2015, pp. 2506–2514.

[20] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, S.-F. Chang, Designing category-level attributes for discriminative visual recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 771–778.

[21] M. Rohrbach, M. Stark, B. Schiele, Evaluating knowledge transfer and zero-shot learning in a large-scale setting, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1641–1648.

[22] Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele, Evaluation of output embeddings for fine-grained image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2927–2936.

[23] Z. Al-Halah, R. Stiefelhagen, How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes, in: IEEE Winter Conference on Applications of Computer Vision, 2015, pp. 837–843.

[24] Z. Al-Halah, M. Tapaswi, R. Stiefelhagen, Recovering the missing link: predicting class-attribute associations for unsupervised zero-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[25] 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, 2013, pp. 2121–2129.

[26] B. Romera-Paredes, P. Torr, An embarrassingly simple approach to zero-shot learning, in: Proceedings of The International Conference on Machine Learning, 2015, pp. 2152–2161.

[27] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, B. Schiele, Latent embeddings for zero-shot classification, in: IEEE International Conference on Computer Vision, 2016.

[28] Z. Akata, M. Malinowski, M. Fritz, B. Schiele, Multi-cue zero-shot learning with strong supervision, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[29] R. Qiao, L. Liu, C. Shen, A. v. d. Hengel, Less is more: zero-shot learning from online textual documents with noise suppression, in: IEEE International Conference on Computer Vision, 2016.

[30] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The Caltech-UCSD Birds-200-2011 Dataset, Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[31] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1778–1785.

[32] G. Patterson, C. Xu, H. Su, J. Hays, The sun attribute database: Beyond categories for deeper scene understanding, International Journal of Computer Vision 108 (2014) 59–81.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (2015) 211–252.

[35] Z. Zhang, V. Saligrama, Zero-shot learning via semantic similarity embedding, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4166–4174.

[36] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[38] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.