A Survey of Zero-Shot Learning on Image classification

Zuoyu Yan
yanzuoyu3@pku.edu.cn
Wangxuan Institute of Computer Technology of Peking University

ABSTRACT

to be continued

KEYWORDS

neural networks, zero-shot learning

ACM Reference Format:
Zuoyu Yan. 2020. A Survey of Zero-Shot Learning on Image classification. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

to be continued

2 TEMPLATE OVERVIEW

3 APPLICATIONS OF ZERO-SHOT LEARNING

3.1 zero-shot learning in image classification

In this section, we mainly classify existing zero-shot learning methods into three categories: (1) methods based on direct semantic prediction (2) methods based on embedding space (3) methods based on generative models. In the following paragraph, each type of work will be discussed in detail.

3.1.1 methods based on direct semantic prediction. Most early zero-sample classification methods can be attributed to methods based on direct semantic prediction[13, 14]. The main idea of such methods is to train a classifier based on semantic attributes, and predict the category of certain images.

The most representative models among these methods are the DAP (direct attribute prediction) and IAP(indirect attribute prediction) proposed by Lampert et al.[13, 14] in 2013. They first used the attribute features between different categories to compute their semantic association. In their work of DAP, training instances from classes that have this attribute constitute the positive instances, and training instances from classes that do not have this attribute constitute the negative instances, with these instances, the classifier for the attributes is learned. After learning the attribute classifiers, the inference from attribute to unseen class is represented by the inference framework in the probability form. While in IAP, instead of direct learning, they first learn classifiers for seen classes, and the classifier for each attribute is the combination of classifiers for seen classes that contain this attribute, the relationship between seen classes and unseen classes are based on their semantic description.

Besides, works such as [6, 10] introduce attribute learning in zero-shot image classification. Fu et al.[6] propose the concept of semi-latent attributes, they first combine the known attributes and the hidden attributes, train a model to learn the half-hidden attributes of the input images, and embed the samples and labels to the half-hidden attributes for KNN(k-nearest neighbor) classification. Jayaraman[10] first uses random forest for zero-shot classification to improve the performance of the model. Although these works have strong interpretability, [9] points out that labels given by human beings are not always reliable, and false labels may lead to the negative performance of these methods. Nonetheless, the relationship between attributes may lead to redundant information, thus lower the performance of these methods[11].

3.1.2 methods based on embedding space. In order to solve the problem mentioned above, researchers propose methods based on embedding space. The main idea of these methods is to reflect the visual features and labels to one certain space, and evaluate their similarity for final prediction. These methods can be categorized as (1) methods based on semantic space. (2) methods based on visual space. (3) methods based on common space.

methods based on semantic space Owing to the advantage that each class is represented by one semantic vector in the semantic space, taking the semantic space as the embedding space is helpful for the better embedded visual data structure. From a different point of view, semantic space can be categorized into euclidean and non-euclidean spaces. Euclidean spaces are more conventional and simpler as the data has a flat representation in such spaces. Yu et al.[27] first propose a model that learns features from the visual distribution of images. Xian et al.[24] hold the idea that most linear models are not suitable for fine-grained image classification, so they construct a piecewise linear model by combining hidden features using compatibility function. [5] is one of the pioneers which introduce deep learning in zero-shot learning, in their paper, they use CNN and Word2Vec features as inputs and use Hinge loss function as their optimization function to construct an end-to-end zero-shot classification model. While the intrinsic relationship between data points is better preserved when the geometrical relation between them is considered, non-euclidean spaces are commonly based on clusters or graph networks. The work produced by Wang et al. [23] processes the unweighted knowledge graph by exploiting recent developments in neural networks for non-Euclidean spaces, such as graph and manifold spaces. [30] proposes a Generalized Zero-Shot learning (GZSL) method that is agnostic to both unseen images and unseen semantic vectors during training. They quantify the impact of noisy semantic data by utilizing a novel visual oracle to visually supervise a learner.

methods based on visual space Despite the fact that methods

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

based on semantic embedding developed dramatically, this strategy will significantly shrink the variance of the data points and thus aggravate the hubness problem(there exist some samples which will become the nearest neighbor of many category prototypes, which may lead to the decrease of performance while using nearest neighbor classification). To deal with the problem, researchers propose methods based on visual space, aiming to learn an embedding function from semantic space to visual space.[21] proposes an effective way to reduce hubness by mapping labels into the example space to suppress the emergence of hubs. Some transductive or semi-supervised Semantic Embedding (SE) methods such as [22] can alleviate the projection domain shift by utilizing the unlabeled test instances from unseen categories in the training phase. SYNC[3] constructs the classifiers of unseen classes by taking the linear combinations of base classifiers, which are trained in a discriminative learning framework. However, using the visual space as the embedding space faces a new problem. Instance features in the visual space are not distributed in an ideal structure due to the possibility of large inter-class similarities and small intra-class similarities.

methods based on common space Based on the ideas above, researchers tried to embed both visual features and semantic features in a common space. Given the data limitations, previous text embedding methods work surprisingly well for zero-shot visual recognition, but there remains a large gap between the text embedding methods and human-annotated attributes [1]. In order to solve this problem, [19] proposes to learn binary mappings between attributes and images mapping by optimizing a simple objective function that has a closed-form solution. In addition, there are also some works based on deep neural networks. Yang et al. [26] and Ba et al. [15] use image and text descriptions as input, respectively extract their features using neural networks and analyze the relationship between these features in a common space, and then use the sorting hinge loss [26] or binary cross-entropy loss [15] to train the network. In addition, Reed et al. [18] firstly uses the image and the corresponding text description as input, while the images are extracted by the CNN model, the text descriptions are extracted by the recurrent neural network. Based on the 0-1 loss function, a cross-modal objective function is established in the output layer, and simultaneously inverted two parts of the network, in order to establish the semantic relationship between different modalities. The work of Zhang et al. [29] is similar to the overall idea of the literature [18], the difference is that the model they proposed can merge text description information with other semantic information.

Although methods based on visual space and common space can alleviate the hubness problem, it can not deal with data imbalance. When training on datasets with few images, the robustness of the model may not be guaranteed.

3.1.3 *methods based on generative model.*

In order to cope with the problem of data imbalance, researchers begin to propose methods based on generative models that aim to learn the probability distribution of the dataset. The main idea of these methods is to use semantic attributes to generate visual samples for training, so as to transform zero-shot learning to supervised learning. With the development of GAN(generative adversarial network), these methods begin to make impressive progress. [17] proposes a conditional Variational Auto-Encoder(cVAE) that learns to generate samples according to given class embeddings. [23] extends this notion with trainable class conditional latent spaces. [12] also develops a cVAE except that their model learns a separate semantic embedding regressor/discriminator. [2] evaluates several generative models for learning to generate training examples. [4] adopts cycle consistency loss of cycle-GAN into zero-shot learning to regularize feature synthesis network. [28] uses a separate reconstructor, discriminator, and classifier all targeting at visual features to remedy the domain-shift problems. Slightly different from mainstream approaches, [16] introduces diffusion regularization to increase the utility of features. [25] proposes a WGAN [8] based formulation that uses a discriminative supervised loss function, in addition to the unsupervised adversarial loss. In this model, the supervised loss enforces the WGAN generator to produce samples that are correctly classified according to a pre-trained classifier of seen classes. Similar to [25], [20] also trains a conditional WGAN towards synthesizing training samples. However, there are two major differences. First, [20] uses the proposed gradient matching loss, which aims to directly maximize the value of the produced training examples by measuring the quality of the gradient signal obtained over the synthesized examples. Second, [20] learns an unconditional discriminator, which does not rely on a semantic embedding vector, aiming to explore the incorporation of unlabeled training examples into training in a semi-supervised fashion. [7] takes class-level semantic embeddings as input, instead of class labels, as the auxiliary information for the sake of capturing a more delicate description for feature generation and make use of intermediate outputs of the neural network for perceptual reconstruction to provide a more semantics-preserving metric for feature generation.

REFERENCES

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. 2015. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2927–2936.

[2] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. 2017. Generating visual representations for zero-shot classification. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2666–2673.

[3] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. 2016. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5327–5336.

[4] Rafael Felix, Vijay BG Kumar, Ian Reid, and Gustavo Carneiro. 2018. Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 21–37.

[5] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*. 2121–2129.

[6] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. 2013. Learning multimodal latent attributes. *IEEE transactions on pattern analysis and machine intelligence* 36, 2 (2013), 303–316.

[7] Rui Gao, Xingsong Hou, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Zhao Zhang, and Ling Shao. 2020. Zero-VAE-GAN: Generating Unseen Features for Generalized and Transductive Zero-Shot Learning. *IEEE Transactions on Image Processing* 29 (2020), 3665–3680.

[8] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Advances in neural information processing systems*. 5767–5777.

[9] Sheng Huang, Mohamed Elhoseiny, Ahmed Elgammal, and Dan Yang. 2015. Learning hypergraph-regularized attribute predictors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 409–417.

[10] Dinesh Jayaraman and Kristen Grauman. 2014. Zero-shot recognition with unreliable attributes. In *Advances in neural information processing systems*. 3464–3472.

[11] Dinesh Jayaraman, Fei Sha, and Kristen Grauman. 2014. Decorrelating semantic visual attributes by resisting the urge to share. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1629–1636.

[12] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. 2018. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4281–4289.

[13] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 951–958.

[14] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence* 36, 3 (2013), 453–465.

[15] Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*. 4247–4255.

[16] Yang Long, Li Liu, Fumin Shen, Ling Shao, and Xuelong Li. 2017. Zero-shot learning using synthesised unseen visual data with diffusion regularisation. *IEEE transactions on pattern analysis and machine intelligence* 40, 10 (2017), 2498–2512.

[17] Ashish Mishra, Shiva Krishna Reddy, Anurag Mittal, and Hema A Murthy. 2018. A generative model for zero shot learning using conditional variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2188–2196.

[18] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 49–58.

[19] Bernardino Romera-Paredes and Philip Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*. 2152–2161.

[20] Mert Bulent Sariyildiz and Ramazan Gokberk Cinbis. 2019. Gradient matching generative networks for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2168–2178.

[21] Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. 2015. Ridge regression, hubness, and zero-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 135–151.

[22] Seyed Mohsen Shojaee and Mahdieh Soleymani Baghshah. 2016. Semi-supervised zero-shot learning by a clustering-based approach. *arXiv preprint arXiv:1605.09016* (2016).

[23] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. 2018. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6857–6866.

[24] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. 2016. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 69–77.

[25] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. 2018. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5542–5551.

[26] Yongxin Yang and Timothy M Hospedales. 2014. A unified perspective on multi-domain and multi-task learning. *arXiv preprint arXiv:1412.7489* (2014).

[27] Felix X Yu, Liangliang Cao, Rogerio S Feris, John R Smith, and Shih-Fu Chang. 2013. Designing category-level attributes for discriminative visual recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 771–778.

[28] Haofeng Zhang, Yang Long, Li Liu, and Ling Shao. 2019. Adversarial unseen visual feature synthesis for zero-shot learning. *Neurocomputing* 329 (2019), 12–20.

[29] Li Zhang, Tao Xiang, and Shaogang Gong. 2017. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2021–2030.

[30] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. 2019. Generalized zero-shot recognition based on visually semantic embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2995–3003.