GSJ: Volume 11, Issue 11, November 2023, Online: ISSN 2320-9186
www.globalscientificjournal.com

Zero-Shot Learning: Methods, Applications & Challenges

Shabana Mir

KeyWords
Deep learning, Few-shot learning, Transfer Learning, Zero-shot classification, Zero-shot learning

ABSTRACT

Zero-shot learning (ZSL) or learning with zero training samples is an innovative machine-learning technique that overcomes the limitations of traditional classification methods. It enables models to learn and classify novel (unseen) classes without using any labeled data from those classes during training. With ZSL, machines can learn about previously unseen concepts, opening up new possibilities in image recognition and text understanding. ZSL could bring about a paradigm shift in various real-world applications, particularly in domains where labeled data is limited or costly such as image classification, natural language processing, and medical diagnosis. This paper provides an easy-to-understand review of ZSL methods, applications, and challenges to facilitate understanding and implementation of zero-shot learning.

1. Introduction

Machine learning models have made remarkable progress in classification tasks but they can only recognize classes they have been trained on and they are typically trained on labeled data. This makes them less useful in real-world situations where there may not be enough labeled data for all classes or getting labeled data is expensive and time-consuming. This is especially true for computer vision and natural language processing tasks that require large and complex datasets. Zero-shot learning addresses this challenge by enabling models to recognize classes they have not seen during training.

Zero-shot learning is a subfield of transfer learning and an extreme case of few-shot learning. It enables a model to learn about unseen classes without using any of its instances during training. It is similar to the way humans learn. Humans can recognize previously unseen objects by utilizing their descriptions. For example, someone who has seen a horse but never a zebra can recognize a zebra if they are told that zebras look like striped horses. Similarly, using seen classes labeled data and semantic information (description) about unseen classes, zero-shot learning enables models to recognize unseen classes.

Compared to recent reviews [1] [2] [3] [4], this work provides an easy-to-understand review of zero-shot learning, including its different methods, data representations, training and testing phases, evaluation techniques, strengths, and limitations to make understanding and implementation of zero-shot learning simple and effective.

2. Data

The data in zero-shot learning contains seen classes, unseen classes, and auxiliary information [5].

- **Seen classes:** It contains existing classes for which labeled images are available. These classes are used during training.
- **Unseen classes:** It contains new classes for which labeled images are not available. The model has no exposure to these classes during training, making it a challenging task to classify them accurately.
- **Auxiliary information:** It contains descriptions, semantic attributes, or word vectors for the seen and unseen classes.

$S = \{(x, y, h_y) \mid x \in X^S , y \in Y^S , h \in A^S\}$
$U = \{(x, y, h_y) \mid x \in X^U , y \in Y^U , h \in A^U\}$
(**S** seen class data, **U** unseen class data, **A** auxiliary info, **X** image set, **Y** class label set, $\mathbf{h_y}$ semantic encoding, **x** image, **y** class label)

Based on the availability of data during the training phase, zero-shot learning is divided into the following categories:

- **Inductive ZSL:** In this type, the training data includes seen classes labeled data and semantic descriptions/attributes for both seen and unseen classes.
- **Transductive ZSL:** In this type, the training data includes seen classes labeled data (from dataset), unseen classes unlabeled data (generated by model), and semantic descriptions/attributes for both seen and unseen classes.

Based on the availability of data during the testing phase, zero-shot learning is divided into the following categories:

- **Conventional ZSL:** In this type, the testing data only includes images from unseen classes.
- **Generalized ZSL:** In this type, the testing data includes both seen and unseen classes during testing.

3. Methods & Models

3.1 Embedding based methods

Embedding-based methods are the most common type of zero-shot learning. They work by learning a shared embedding space (visual, semantic, or hybrid) for both seen and unseen classes. This method requires only seen class data during training. During training, a deep model learns to map visual space (image vector) to semantic space (word vector) using data from seen classes. During testing, the unseen class image vector is passed as input to the trained network which outputs the corresponding word vector. For classification, a nearest neighbor search is performed in the semantic space to find the closest match to the output of the network. The drawback of this approach is that it can be biased because the model is trained only on seen classes.

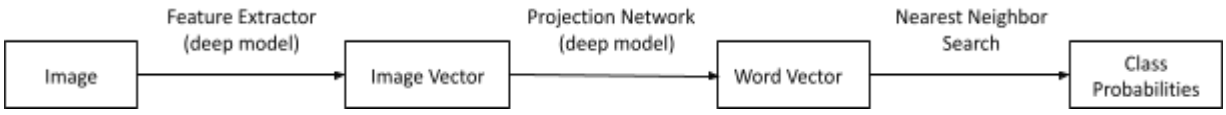

Figure 1: Basic steps of zero-shot learning using embedding approach

3.2 Generative model based methods

Generative model-based methods are an advanced approach to zero-shot learning that avoids the bias and domain shift issues of embedding methods. This is done by using both seen and unseen class data during training. As unseen class data is limited, a generative model is used to generate the unseen class data using semantic attributes. A conditional generative adversarial network is trained to map semantic space (word vector) to visual space (image vector). Once the unseen class image features have been generated, a simple classifier is trained on the seen and unseen class image vectors to learn and classify the data.

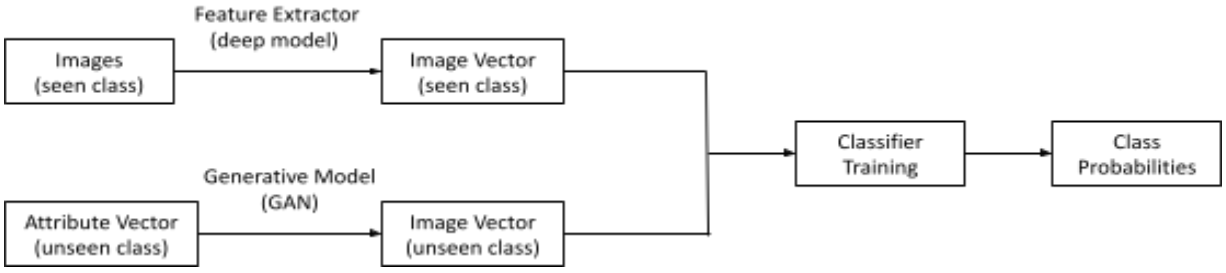

Figure 2: Basic steps of zero-shot learning using generative approach

3.3 Evaluation

The evaluation of zero-shot models is different from traditional classification methods. To evaluate zero-shot model performance, average per-class-top-1 accuracy (a_Y) and harmonic mean (H) are used [5].

In average per-class-top-1 accuracy (a_Y), we first find the accuracy of each class separately and then average it over all classes.

$$a_Y = \frac{1}{n} \sum_{c=1}^{n} \frac{no.\ of\ correct\ predictions\ in\ c}{total\ predictions\ in\ c}$$ *(**Y** set of classes, **c** class, **n** total classes in set Y)*

Harmonic mean is used in the case of Generalized ZSL settings. We use the top-1 accuracies for seen and unseen classes to compute the harmonic mean.

$$H = \frac{2 \times a_{Y^u} \times a_{Y^s}}{a_{Y^u} + a_{Y^s}}$$ *(**U** unseen class, **S** seen class)*

3.4 Zero-shot algorithms and pretrained models

The state-of-the-art zero-shot algorithms are Semantic Output Codes (SOC) [6], Convex Combination of Semantic Embedding (ConSE) [7], Embarrassingly Simple Zero Shot Learning (ESZSL) [8], Structured Joint Embeddings (SJE) [9], SynC Synthesized Classifiers (SyncC) [10], Latent Embeddings (LatEM) [11], and Missing Data Problem (MDP) [12].

CLIP (Contrastive Language-Image Pre-Training) is a pre-trained zero-shot classifier developed by OpenAI [13]. Given an image and text description, the model can predict the relevant text description for that image. Other pre-trained models are ALIGN, CLIPSeg, Chinese-CLIP, AltCLIP, X-CLIP, VisualBERT, BLIP, LXMERT [14].

4. Applications

Zero-shot learning has a wide range of applications, especially in computer vision and natural language processing. It can be used in:

- **Image classification** to classify unseen images e.g visual search engines
- **Object detection** to detect novel objects e.g autonomous vehicles [2]
- **Text classification** to classify text into topics e.g unseen emotion recognition [15]
- **Text Translation** e.g translation systems [16]
- **Semantic segmentation** to segment unseen object categories e.g COVID x-ray diagnosis [2]
- **Resolution Enhancement** to enhance image resolution without predefined high-resolution images e.g single-image super-resolution [17]
- **Audio Processing** e.g zero-shot based voice conversion [18]
- **Image Retrieval** e.g sketch-based image retrieval [19]
- **Image Generation** e.g text/sketch-to-image generation [20]

- **Action Recognition** e.g human-object interaction recognition [21]
- **Style Transfer** e.g artistic style transfer [22]

5. Challenges

In this section, we discuss the issues in zero-shot learning that affect the model performance and possible solutions to address these challenges. The most common challenges in zero-shot classification are bias, hubness, and domain shift.

In inductive training settings, zero-shot models are trained on seen classes, so they are biased toward predicting seen classes at test time. This can be a problem when the model is tested on images from seen and unseen classes. A solution to the bias problem is transductive learning.

Hubness happens when high-dimensional vectors are projected into low-dimensional spaces, and the projected points are clustered around a few hubs because such projection reduces variance. This can make it difficult for the model to accurately classify unseen classes. Visual embedding space can mitigate the hubness problem because visual space better preserves the structure.

Domain shift occurs when the training and testing data come from different distributions. This happens in zero-shot learning because the seen classes on which the model is trained are different from the unseen classes on which it is tested. Transductive setting can overcome the domain-shift issue.

6. Conclusion

This paper provides a comprehensive and easy-to-understand overview of zero-shot learning to help readers better understand and implement this powerful technique. We discussed the fundamentals of zero-shot learning, including data representation and preparation, training and testing phases, zero-shot model categories, methods, applications, and challenges.

References

[1] W. Wang, V. W. Zheng, H. Yu, and C. Miao, “A survey of zero-shot learning: Settings, methods, and applications”, *ACM Transactions on Intelligent Systems and Technology*, 2019, doi: 10.1145/3293318
[2] M. Rezaei, M. Shahidi, “Zero-shot learning and its applications from autonomous vehicles to COVID-19 diagnosis: A review”, *Intelligence-Based Medicine*, 2020, doi: 10.1016/j.ibmed.2020.100005
[3] G. Yang, Z. Ye, R. Zhang, and K. Huang, “A comprehensive survey of zero-shot image classification: methods, implementation, and fair evaluation”. *Applied Computing and Intelligence*. 2022, doi: 10.3934/aci.2022001
[4] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, “Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly” 2020, available at https://arxiv.org/pdf/1707.00600.pdf
[5] S. Chandhok, "Zero-shot Learning : An Introduction," LearnOpenCV, 2020, available at https://learnopencv.com/zero-shot-learning-an-introduction/
[6] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, “Zero-shot Learning with Semantic Output Codes”, 2014, available at https://papers.nips.cc/paper_files/paper/2009/file/1543843a4723ed2ab08e18053ae6dc5b-Paper.pdf
[7] 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 ”, 2014, available at https://arxiv.org/pdf/1312.5650.pdf
[8] F. Minhas, “Python Implementation of Embarrassingly Simple Zero Shot Learning”, 2015, available at https://github.com/foxtrotmike/ESZSL
[9] Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele, “Evaluation of Output Embeddings for Fine-Grained Image Classification”, 2015, doi: 10.48550/arXiv.1409.8403
[10] S. Changpinyo, W. Chao, B. Gong, F. Sha, “Synthesized Classifiers for Zero-Shot Learning”, 2016, doi: 10.48550/arXiv.1603.00550
[11] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, “Latent Embeddings for Zero-shot Classification”, 2016, doi: 10.48550/arXiv.1603.08895
[12] B. Zhao, B. Wu, T. Wu, and Y. Wang, “Zero-Shot Learning posed as a Missing Data Problem”, 2017, doi: 10.48550/arXiv.1612.00560
[13] OpenAI, "CLIP: Connecting Text and Images," 2021, available at https://openai.com/blog/clip/
[14] Hugging Face, “Multimodal Models”, available at https://huggingface.co/docs/transformers/model_doc
[15] C. Zhan, D. She, S. Zhao, M.M. Cheng, and J. Yang “Zero-shot emotion recognition via affective structural embedding”, 2019, doi: 10.1109/ICCV.2019.00124
[16] J. Gu, Y. Wang, K. Cho, and V.O.K. Li, “Improved zero-shot neural machine translation via ignoring spurious correlations”, 2019, doi: 10.18653/v1/P19-1121
[17] A. Shocher, N. Cohen, and M. Irani “Zero-shot super-resolution using deep internal learning”, 2018, doi: 10.1109/CVPR.2018.00329
[18] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson “Autovc: zero-shot voice style transfer with only autoencoder loss”, 2019
[19] Y. Long, L. Liu, Y. Shen, and L. Shao, “Towards affordable semantic searching: zero-shot retrieval via dominant attributes”, 2018
[20] S. K. Yelamarthi, S. K. Reddy, A. Mishra, and A. Mittal, “A Zero-Shot Framework for Sketch Based Image Retrieval”, 2018
[21] J. Gao, T. Zhang, and C. Xu, “I know the relationships: zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs”, 2019, doi: 10.1609/aaai.v33i01.33018303
[22] L. Sheng, Z. Lin, J. Shao, and X. Wang, “Avatar-net: multi-scale zero-shot style transfer by feature decoration”, 2018, doi: 10.1109/CVPR.2018.00860