2021 IEEE International Conference on Image Processing

Generalized Zero-Shot Recognition through Image-Guided Semantic Classification

Fang Li and Mei-Chen Yeh

Department of Computer Science and Information Engineering, National Taiwan Normal University

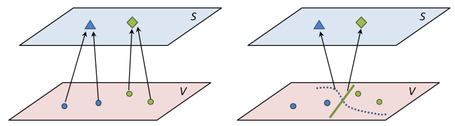
Introduction

Zero-shot learning (ZSL) aims to recognize objects whose instances have not been seen during training.

A majority of ZSL methods can be viewed using the **visual-semantic embedding framework**.

Images are mapped from the visual space to the semantic space in which all classes reside, or images and labels are projected to a latent space. Then, the inference is performed in this common space, typically using cosine similarity or Euclidean distance.

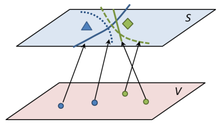
Another perspective of embedding-based methods is to construct an image classifier for each unseen class by learning the correspondence between a binary one-versus-rest image classifier and its class prototype in the semantic space.



Contributions

We propose IGSC, which aims to learn the correspondence between an image and its corresponding label classifier.

- IGSC analyzes the input image and seeks for combinations of variables in the semantic space (e.g., combinations of attributes) that distinguish a class (belonging to the input) from other classes.
- IGSC learns the correspondence between an image in the visual space and a classifier in the semantic space.
- The correspondence can be learned with training pairs in the scale of training images rather than that of classes.
- Label classification is conducted by a semantic classifier whose weights are generated on the fly.



Method

Problem Definition

Given a training set $S = \{(x_n, y_n), n = 1..N\}$, with $y_n \in Y_S$ being a class label in the seen class set, the goal of GZSL is to learn a classifier $f: X \longrightarrow Y$ which can generalize to predict any image x to its correct label, which is not only in Y_S but also in the unseen class set Y_U.

Image-Guided Semantic Classification Model

The compatibility function is achieved by implementing two functions, $h(\theta(x), W)$ and $g(\varphi(y), M)$, as illustrated in the figure below.

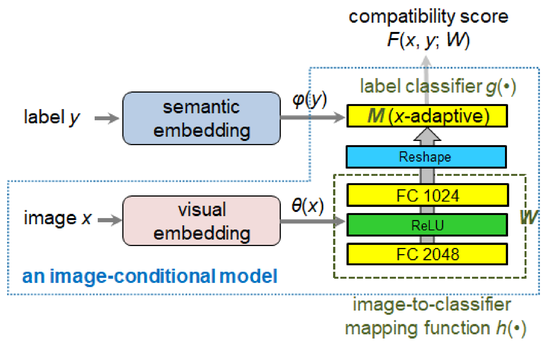
The first function $h(\cdot)$ receives an image embedding as input and returns parameters *M* characterizing a label classifier: $M = h(\theta(x), W)$.

The second function $g(\cdot)$ is a label classifier, characterized by the parameters outputted by $h(\cdot)$. This function takes a label vector as input, and returns a prediction score indicating the probability of the label belonging to the input image: $s = g(\varphi(y), M)$.

The final compatibility score is obtained by normalizing the prediction scores to probabilistic values with softmax:

$$F(x, y_j; W) = \frac{\exp(s_j)}{\sum_k \exp(s_k)}.$$

The model parameters *W* are learned by minimizing the cross entropy loss.



The architecture of IGSC. This model receives an image and a label, and it returns the compatibility score of this input pair. The score indicates the probability of the label belonging to the image. The score is calculated by a label classifier g(·), whose weights M are stored in the output layer of a fully connected neural network. Therefore, weight values depend on the input image. The neural network is characterized by the parameters W, which are the only parameters required to learn from training data.

The IGSC method has the following characteristics:

- IGSC learns the correspondence between an image in the visual space and a classifier in the semantic space. The correspondence can be learned with training pairs in the scale of training images rather than that of classes.
- IGSC performs learning to learn in an end-to-end manner. Label classification is conducted by an image-conditioned semantic classifier whose weights are generated on the fly. This model is simple yet powerful because of its adaptive nature.
- IGSC unifies visual attribute detection and label classification. This is achieved via the design of a conditional network (the proposed classifier learning method), in which label classification is the main task of interest and the conditional input image provides additional information of a specific situation.
- IGSC alleviates the hubness problem. The correspondence between an image and a semantic classifier learned from seen classes can be transferred to recognize unseen concepts.

Results

We used four popular benchmark datasets—including SUN, CUB, AWA2, and aPY—for evaluating the proposed method.

We followed the standard evaluation metrics used in the literature. We reported acc_s (test images are from seen classes and the prediction labels are the union of seen and unseen classes), acc_u (test images are from unseen classes and the prediction labels are the union of seen and unseen classes), and the harmonic mean.

	SUN			CUB			AWA2			aPY		
Method	acc_u	acc_s	H	acc_u	acc_s	H	acc_u	acc_s	H	acc_u	acc_s	H
LATEM [9]	14.7	28.8	19.5	15.2	57.3	24.0	11.5	77.3	20.0	1.3	71.4	2.6
DEVISE [4]	16.9	27.4	20.9	23.8	53.0	32.8	17.1	74.7	27.8	3.5	**78.4**	6.7
ESZSL [8]	11.0	27.9	15.8	14.7	56.5	23.3	5.9	77.8	11.0	2.4	70.1	4.6
SYNC [20]	7.9	**43.3**	13.4	11.5	70.9	19.8	9.7	89.7	17.5	7.4	66.3	13.3
SP-AEN [18]	24.9	38.6	30.3	34.7	70.6	46.6	23.3	90.9	37.1	13.7	63.4	22.6
PSR [13]	20.8	37.2	26.7	24.6	54.3	33.9	20.7	73.8	32.3	13.5	51.4	21.4
DCN [6]	25.5	37.0	30.2	28.4	60.7	38.7	–	–	–	14.2	75.0	23.9
AREN [21]	19.0	38.8	25.5	38.9	**78.7**	**52.1**	5.6	**92.9**	26.7	9.2	76.9	16.4
DAZLE [22]	21.7	31.9	25.8	**42.0**	65.3	51.1	**25.7**	82.5	39.2	–	–	–
IGSC	**39.4**	31.3	**34.9**	40.8	60.2	48.7	**25.7**	83.6	**39.3**	**23.1**	58.9	**33.2**

Generalized zero-shot learning results (top-1 accuracy and H) on four benchmark datasets. All methods are agnostic to both unseen images and unseen semantic vectors during training.

All methods under comparison—including the proposed method—are inductive to both unseen images and unseen semantic vectors.

By examining the harmonic mean values, IGSC consistently outperformed other competitive methods on three out of the four datasets. The performance gain validated the effectiveness of learning image guided semantic classifiers.

Compared with embedding based methods, this novel paradigm not only has more training pairs (in the scale of the training images) for learning the correspondence between an image and its corresponding label classifier but also allows different manners to separate classes based on the content of input image.

Compared with attribute based methods which take a two-step pipeline to detect attributes from one image and aggregate the detection results for label prediction, IGSC unifies the steps.

Compared with recent methods, IGSC is much simpler and therefore has a greater flexibility.

Conclusion

We propose a visual-semantic embedding model that transforms an image into a label classifier, consequently used to predict the correct label in the semantic space.

Modeling the correspondence between an image and its label classifier enables a powerful method that achieves promising performances on four benchmark datasets.

Contact:

Mei-Chen Yeh
Professor, National Taiwan Normal University

Email: myeh@ntnu.edu.tw
Web: http://www.csie.ntnu.edu.tw/~myeh

Acknowledgements

This work was supported by the Ministry of Science and Technology of Taiwan (MOST 108-2221-E-003-017-MY2) and Qualcomm (NAT-414697).