Zero-Shot Object Recognition by Semantic Manifold Distance

Zhenyong Fu, Tao Xiang, Elyor Kodirov, Shaogang Gong
School of Electronic Engineering and Computer Science, Queen Mary University of London.

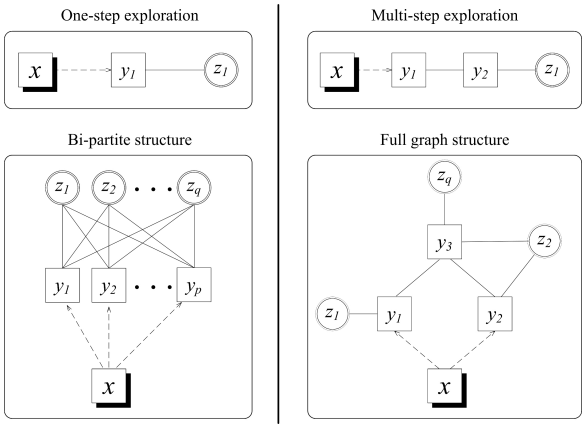

Figure 1: One-step exploration vs multi-step exploration to the semantic relationship. One-step exploration leads to a bi-partite structure, while multi-step leads to a full graph structure for the semantic relationship.

Object recognition by zero-shot learning (ZSL) aims to recognise objects without seeing any visual examples by learning knowledge transfer between seen and unseen object classes. Existing works usually utilize the one-step exploration to the semantic relationship [1, 2], while in this paper we consider to exploit the multi-step exploration. As shown in Fig. 1, one-step exploration natually leads to a bi-partite structure, while multi-step exploration leads to a full graph structure for the semantic relationship. The rich intrinsic structure of the semantic categories is modelled by a semantic (label) graph. Subsequently, the conventional distance metric (e.g. cosine) is replaced by the distance along the semantic manifold, which is computed through an absorbing Markov chain process (AMP).

Let $\mathcal{Y} = \{y_1, \ldots, y_p\}$ denotes a set of p seen class labels and $\mathcal{Z} = \{z_1, \ldots, z_q\}$ a set of q unseen class labels. These two sets of labels are disjoint, i.e. $\mathcal{Y} \cap \mathcal{Z} = \varnothing$. We are given a labelled training dataset $X_{\mathcal{Y}} = \{(\mathbf{x}_j, y_j)\}$ where $\mathbf{x}_j$ is a d-dimensional feature vector extracted from the j-th labelled image and $y_j \in \mathcal{Y}$. In addition, a test dataset $X_{\mathcal{Z}} = \{(\mathbf{x}_i, y_i)\}$ is provided where $\mathbf{x}_i$ is a d-dimensional feature vector extracted from the i-th unlabelled test image and the unknown $y_i \in \mathcal{Z}$. The goal of ZSL is to learn a classifier $f : X \to \mathcal{Z}$ to predict y_i.

We define an absorbing Markov chain process on the semantic graph as follows. Each unseen class node is viewed as an *absorbing* state and each seen class node is viewed as a *transient* state, whilst the transition probability from class node i to class node j is $p_{ij} = w_{ij} / \sum_j w_{ij}$. An absorbing state means that for each unseen class node i, we have $p_{ii} = 1$ and $p_{ij} = 0$ for $i \neq j$. We re-number the class nodes (as states in a Markov process) so that the seen class nodes (transient states) come first. Then, the transition matrix P of the above absorbing Markov chain process has the following canonical form:

$$P = \left(\begin{array}{c|c} Q_{p\times p} & R_{p\times q} \\ \hline \mathbf{0}_{q\times p} & \mathrm{I}_{q\times q} \end{array} \right). \tag{1}$$

In Eq. (1), $Q_{p\times p}$ describes the probability of transitioning from a transient state (seen class) to another, $R_{p\times q}$ describes the probability of transitioning from a transient state (seen class) to an absorbing state (unseen class). In addition, $\mathbf{0}_{q\times p}$ and the identity matrix $\mathrm{I}_{q\times q}$ denote that the absorbing Markov chain process cannot leave the absorbing states once it arrives.

For ZSL, we first incorporate $\mathbf{x}_i$ into the semantic graph. This is followed by applying an extended absorbing Markov chain process (see Fig. 2). For $\mathbf{x}_i$, we have $T_i = [t_{ij}]_{1\times p}$ as a row vector of p elements. Each element is $t_{ij} = p(y_j|\mathbf{x}_i)$. Each test image $\mathbf{x}_i$ is incorporated into the semantic graph as

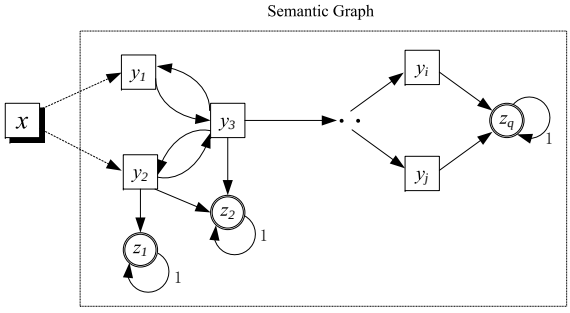

Figure 2: After incorporating a test image into a semantic graph, zero-shot learning can be viewed as an extended absorbing Markov chain process (AMP) on the graph.

a transient state. The transition matrix $\tilde{P}$ of the extended absorbing Markov chain process have the following canonical form:

$$\tilde{P} = \left(\begin{array}{cc|c} Q_{p\times p} & \mathbf{0}_{p\times 1} & R_{p\times q} \\ (T_i)_{1\times p} & 0_{1\times 1} & \mathbf{0}_{1\times q} \\ \hline \multicolumn{2}{c|}{\mathbf{0}_{q\times(p+1)}} & \mathrm{I}_{q\times q} \end{array} \right). \tag{2}$$

Formally, the absorbing probability b_{ij} is the probability that the absorbing Markov chain will be absorbed in the absorbing state s_j if it starts from the transient state s_i. The absorbing probability matrix $\tilde{B} = [b_{ij}]_{(p+1)\times q}$ can be computed as follows:

$$\tilde{B} = \tilde{N} \times \tilde{R}, \tag{3}$$

in which $\tilde{N}$ is the fundamental matrix of the extended absorbing Markov chain process and its last row can be computed as:

$$\tilde{N}_{(p+1),\cdot} = \left(\begin{array}{cc} (T_i)(I-Q)^{-1}, & 1 \end{array} \right)_{1\times(p+1)} \tag{4}$$

and then we further compute $\tilde{B}_{p+1,\cdot}$ as

$$\tilde{B}_{p+1,\cdot} = (\tilde{N}_{(p+1),\cdot}) \times \tilde{R} = T_i \times (I-Q)^{-1} R, \tag{5}$$

which can be viewed as the *semantic manifold distance* between the test image and each unseen class.

For the whole test dataset with n images, we use a matrix $S_{n\times q}$ to store the computed absorbing probabilities, in which the i-th row $S_{i,\cdot}$ of S equals to the absorbing probabilities of $\mathbf{x}_i$. If we stack the results of all test images together, we have the final matrix S as follows:

$$S = T(I-Q)^{-1}R. \tag{6}$$

For the test image $\mathbf{x}_i$, we assign it to the unseen label that has the maximum absorbing probability when the absorbing chain starts from $\mathbf{x}_i$. Finally, our ZSL classifer is

$$f(\mathbf{x}_i) = \arg\max_{z_j} S_{i,j} \tag{7}$$

[1] C Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot learning of object categories. 2013.

[2] Marcus Rohrbach, Michael Stark, György Szarvas, Iryna Gurevych, and Bernt Schiele. What helps where–and why? semantic relatedness for knowledge transfer. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 910–917. IEEE, 2010.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.