Attribute Embedding with Visual-Semantic Ambiguity Removal for Zero-shot Learning

Yang Long[1]
ylong2@sheffield.ac.uk
Li Liu[2]
li2.liu@northumbria.ac.uk
Ling Shao[2]
ling.shao@ieee.org

[1] Department of Electronic and Electrical Engineering
The University of Sheffield
Sheffield , UK

[2] Department of Computer and Information Sciences
Northumbria University
Newcastle upon Tyne, UK

Conventional *zero-shot learning* (ZSL) methods recognise an unseen instance by projecting its visual features to a semantic space that is shared by both seen and unseen categories [1, 2]. However, we observe that such a one-way paradigm suffers from the *visual-semantic ambiguity* problem. As shown in Fig. 1, the semantic concepts (e.g. attributes or classes) cannot explicitly correspond to visual patterns, and similar visual features may come from different classes. Such a problem can lead to a huge variance in the visual features for each attribute.

In this paper, we propose the ***V**isual-**S**emantic **A**mbiguity **R**emoval* (VSAR) algorithm to address such a problem. In particular, we propose a novel latent attribute space $\mathcal{V}$ to mitigate the gap between visual and semantic spaces $\mathcal{X}$ and $\mathcal{A}$:

$$J = \|\mathcal{X} - U_1\mathcal{V}\|_F^2 + \alpha\|\mathcal{A} - U_2\mathcal{V}\|_F^2 + \lambda\mathcal{R}, \quad (1)$$

where U_1 and U_2 are two projection matrices. $\mathcal{R}$ is a *Dual-graph* regularisation that combines two supervised graphs $W_{\mathcal{X}}$ and $W_{\mathcal{A}}$ that model the intrinsic data structures in $\mathcal{X}$ and $\mathcal{A}$. In the embedding space $\mathcal{V}$, we expect that if the vertices in both graphs are connected, each pair of embedded points v_i and v_j are also closed to each other. However, for the *visual-semantic ambiguity* problem, $W_{\mathcal{X}}$ and $W_{\mathcal{A}}$ usually give contradictory results. To compromise such conflict, we linearly combine the two graphs, i.e. $W_{ij} = W_{\mathcal{X}_{ij}} + \alpha W_{\mathcal{A}_{ij}}$. The resulted regularisation is:

$$\mathcal{R} = \frac{1}{2}\sum_{i,j=1}^{N} \|v_i - v_j\|^2 W_{ij} = Tr(\mathcal{V}L\mathcal{V}^T), \quad (2)$$

where D is the degree matrix of W, $D_{ii} = \sum_i w_{ij}$. L is known as graph Laplacian matrix $L = D - W$

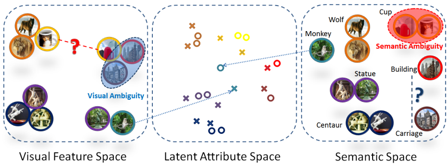

Figure 1: *Visual Ambiguity* (in blue oval): the image of a carriage is taken with a building background. It cannot recover the semantic distance (blue question mark) to the building category. *Semantic Ambiguity* (in red oval): the cup printed with a wolf and the cup-like building share the same name which can lead to a large visual variance (the red question mark). After embedding to the latent attribute space using VSAR, such ambiguity is mitigated.

and $Tr(.)$ computes the trace of a matrix.

Once we obtain the latent attribute embedding $\mathcal{V}$ of the seen data, performing zero-shot recognition is straightforward via *least-square approximation* between $\mathcal{V}$ and $\{\mathcal{A}, \mathcal{X}\}$. During the test, given unseen category names and their attributes in pairs: $\{\mathcal{Y}_u, \mathcal{A}_u\}$. We firstly embed all unseen attributes $\mathcal{A}_u$ into the latent embedding space as references: $\mathcal{V}_u = \mathcal{V}\mathcal{A}^T(\mathcal{A}\mathcal{A}^T)^{-1}\mathcal{A}_u$. Given a test unseen instance $\hat{x}$, its embedded latent attribute representation is: $\hat{v} = \mathcal{V}\mathcal{X}^T(\mathcal{X}\mathcal{X}^T)^{-1}\hat{x}$. Finally, we adopt a simple NN classifier to predict the category label $\hat{c}$:

$$\hat{c} = \arg\min_c \|\hat{v} - v_c\|^2, \text{ where } v_c \in \mathcal{V}_u. \quad (3)$$

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013.

[2] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.