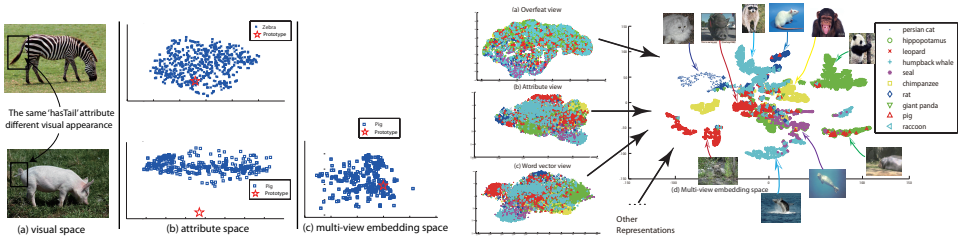
Transductive Multi-class and Multi-label Zero-shot Learning

Yanwei Fu, Yongxin Yang, Timothy M. Hospedales,
Tao Xiang, Shaogang Gong

School of EECS, Queen Mary University of London, UK
Email:{y.fu,yongxin.yang, t.hospedales, t.xiang, s.gong}@qmul.ac.uk



(A) zero-shot prototypes: red stars; image feature projections: blue

(B) (a) overfeat view (b) attribute view (c) word vector view (d) multi-view embedding space

Figure 1. t-SNE visualisation of AwA: (A) Projection domain shift problem; (B) Embedding spaces.

Recently, zero-shot learning (ZSL) has received increasing interest. The key idea underpinning existing ZSL approaches is to exploit knowledge transfer via an intermediate-level semantic representation which is assumed to be shared between the auxiliary/source dataset and the target/test dataset and re-used as a bridge between the source and target domains for knowledge transfer. The semantic representation used in existing approaches varies from visual attributes [10,2,12,6] to semantic word vectors [3,18] and semantic relatedness [16]. However, the overall pipeline is similar: a projection function mapping low-level features to the semantic representation is learned from the auxiliary dataset by either classification or regression models and applied directly to map each instance into the same semantic representation space where a zero-shot classifier is used to recognise the unseen target class instances with a single known 'prototype' of each target class.

1 Transductive multi-class zero-shot learning

Two inherent problems exist in this multi-class ZSL approach. (1) projection domain shift problem: Since the two datasets have different and potentially unrelated classes, the underlying data distributions of the classes differ, so do the 'ideal' projection functions between the low-level feature space and the semantic

spaces. Therefore, using the projection functions learned from the auxiliary dataset/domain without any adaptation to the target dataset/domain causes an unknown shift/bias. This is illustrated in Fig. 1(A), both of Zebra (auxiliary) and Pig (target) classes in AwA dataset share the same 'hasTail' semantic attribute, yet with different visual appearance of their tails. Similarly, many other attributes of Pig are visually different from the corresponding attributes in the auxiliary classes. Figure 1(A-b) illustrates the projection domain shift problem by plotting an 85D attribute space representation of image feature projections and class prototypes: a large discrepancy between the Pig prototype and the projections of its class member instances, but not for Zebra. Such a discrepancy inherently degrades the effectiveness of ZSL of Pig class. To our knowledge, this problem has neither been identified nor addressed in the zero-shot learning literature. (2) *Prototype sparsity problem*: for each target class, we only have a single prototype which is insufficient to fully represent what that class looks like. As shown in Figs. 1(B-b) and (B-c), there often exists large intra-class variations and inter-class similarities. Consequently, even if the single prototype is centred among its class members in the semantic representation space, existing ZSL classifiers still struggle to assign the correct class labels to these highly overlapped data points – one prototype per class simply is not enough to represent the intra-class variability. This problem has never been explicitly identified although a partial solution exists [15].

In addition to these inherent problems, conventional approaches to ZSL are also limited in **exploiting multiple intermediate semantic spaces/views**, each of which may contain complementary information – they are useful in distinguishing different classes in different ways. In particular, while both visual attributes [10,2,12,6] and linguistic semantic representations such as word vectors [13,3,18] have been independently exploited successfully, it remains unattempted and not straightforward to exploit synergistically multiple semantic 'views'. This is because they are often of very different dimensions and types and each suffers from different domain shift effects discussed above. This exploitation has to be transductive for zero-shot learning as only unlabelled data are available for the target classes and the labelled auxiliary data cannot be used directly due to the projection domain shift problem.

In our work [7,5], we propose to solve the projection domain shift problem using a transductive multi-view embedding framework. Under our framework, each unlabelled instance from the target classes is represented by multiple views: its low-level feature view and its (biased) projections in multiple semantic spaces (visual attribute space and word space in this work). We introduce a multi-view semantic space alignment process to correlate different semantic views and the low-level feature view by projecting them onto a latent embedding space learned using multi-view Canonical Correlation Analysis (CCA) [9]. The objective of learning this new embedding space is to transductively (using the unlabelled target data) align the semantic views with each other, and with the low-level feature view to rectify the projection domain shift and exploit their complementarity. Even with the proposed transductive multi-view embedding framework, the pro-

totype sparsity problem remains – instead of one prototype per class, a handful are now available depending on how many views are embedded, which are still sparse. Our solution to this problem is to explore the manifold structure of the data distributions of different views projected onto the same embedding space via label propagation on a graph. To this end, we introduce novel transductive multi-view Bayesian label propagation (TMV-BLP) algorithm for recognition in [5] which combines multiple graphs by Bayesian model averaging in the embedding space. In our journal version [7], we further introduce a novel transductive multi-view hypergraph label propagation (TMV-HLP) algorithm for recognition. The core in our TMV-HLP algorithm is a new distributed representation of graph structure termed heterogeneous hypergraph – instead of constructing hypergraphs independently in different views (i.e. homogeneous hypergraphs), data points in different views are combined to compute multi-view heterogeneous hypergraphs. This allows us to exploit the complementarity of different semantic and low-level feature views, as well as the manifold structure of the target data to compensate for the impoverished supervision available in the form of the sparse prototypes. Zero-shot learning is then performed by semi-supervised label propagation from the prototypes to the target data points within and across the graphs. Some results are shown in Tab. 1 and Fig. 1(B).

Approach	AwA ($\mathcal{H}$ [10])	AwA ($\mathcal{O}$)	AwA ($\mathcal{O},\mathcal{D}$)	USAA	CUB ($\mathcal{O}$)	CUB ($\mathcal{F}$)
DAP	40.5([10]) / 41.4([11]) / 38.4*	51.0*	57.1*	33.2([6,4]) / 35.2*	26.2*	9.1*
IAP	27.8([10]) / 42.2([11])	–	–	–	–	–
M2LATM [6]	41.3	–	–	41.9	–	–
ALE/HLE/AHLE [1]	37.4/39.0/43.5	–	–	–	–	18.0
Mo/Ma/O/D [17]	27.0 / 23.6 / 33.0 / 35.7	–	–	–	–	–
PST [15]	42.7	54.1*	62.9*	36.2*	38.3*	13.2*
[19]	43.4	–	–	–	–	–
[20]	48.3**	–	–	–	–	–
TMV-BLP[5]	47.1	–	–	47.8	–	–
TMV-HLP [7]	**49.0**	**73.5**	**80.5**	**50.4**	**47.9**	**19.5**

Table 1. Comparison with the state-of-the-art on zero-shot learning on AwA, USAA and CUB. Features $\mathcal{H}$, $\mathcal{O}$ and $\mathcal{F}$ represent hand-crafted, OverFeat and Fisher Vector respectively. Mo, Ma, O and D represent the highest results in the mined object class-attribute associations, mined attributes, objectness as attributes and direct similarity methods used in [17] respectively. '–': no result reported. *: our implementation. **: requires additional human interventions.

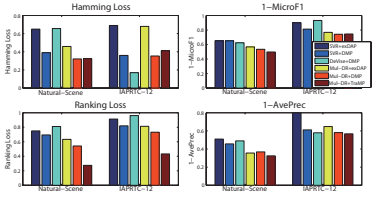
2 Transductive multi-label zero-shot learning

Many real-world data are intrinsically multi-label. For example, an image on Flickr often contains multiple objects with cluttered background, thus requiring more than one label to describe its content. And different labels are often correlated (e.g. cows often appear on grass). In order to better predict these labels given an image, the label correlation must be modelled: for n labels, there are 2^n possible multi-label combinations and to collect sufficient training samples for each combination to learn the correlations of labels is infeasible. More fundamentally, existing multi-class ZSL algorithms cannot model any such correlation as no labeled examples are available in this setting.

We propose a novel framework for multi-label zero-shot learning [8]. Given an auxiliary dataset containing labelled images, and a target dataset *multi-labelled* with unseen classes (i.e. none of the labels appear in the training set),

we aim to learn a zero-shot model that performs multi-label classification on the test set with unseen labels. Zero-shot transfer is achieved using an intermediate semantic representation in the form of the skip-gram word vectors [14] which allows vector-oriented reasoning. For example, $Vec(`Moscow')$ is closer to $Vec(`Russia') + Vec(`capital')$ than $Vec(`Russia')$ or $Vec(`capital')$ only. This property will enable zero-shot multi-label prediction by enabling synthesis of multi-label prototypes in the semantic word space.

Our framework has two main components: multi-output deep regression (Mul-DR) and zero-shot multi-label prediction (ZS-MLP). Mul-DR is a 9 layer neural network that exploits convolutional neural network (CNN) layers, and includes two multi-output regression layers as the final layers. It learns from auxiliary data the mapping from raw image pixels to a linguistic representation defined by the skip-gram language model [14]. With **Mul-DR**, each test image is now projected into the semantic word space where the unseen labels and their combinations can be represented as data points without the need to collect any visual data. **ZS-MLP** addresses the multi-label ZSL problem in this semantic word space by exploiting the property that label combinations can be synthesised. We exhaustively synthesise the power set of all possible prototypes (i.e., combinations of multi-labels) to be treated as if they were a set of labelled instances in the space. With this synthetic dataset, we are able to propose two new multi-label algorithms – direct multi-label zero-shot prediction (DMP) and transductive multi-label zero-shot prediction (TraMP). However, Mul-DR is learned using the auxiliary classes/labels, so it may not generalise well to the unseen classes/labels (projection domain shift problem, as discussed in the previous section). To overcome this problem, we further exploit self-training to adapt Mul-DR to the test classes to improve its generalisation capability. The experimental results on Natural Scene and IAPRTC-12 in Fig 2 show the efficacy of our framework for multi-label ZSL over a variety of baselines. For more details, please read our paper [8].



(A)

Groundtruth	sand−beach, mountain, sky	landscape−nature, mountain, sky	grass	sand−beach, sky
Mul−DR+DMP	sand−beach, sky	landscape−nature, mountain, sky	grass	sand−beach, sky
Mul−DR+TraMP	sand−beach, mountain, sky	landscape−nature, mountain, sky	grass, ground, landscape−nature	ground, sky, sand−beach
DeViSE+DMP	sky	−	−	sky

(B)

Figure 2. (A) Comparing different zero-shot multi-label classification methods on Natural Scene and IAPRTC-12.So smaller values for all metrics are preferred. (B) Examples of ML-ZSL predictions on IAPRTC-12. Top 8 most frequent labels of landscape-nature branch are considered.

References

1. Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013.
2. A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
3. A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model andrea. In *NIPS*, 2013.
4. Y. Fu, T. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *European Conference on Computer Vision*, 2012.
5. Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *ECCV*, 2014.
6. Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning multi-modal latent attributes. *TPAMI*, 2013.
7. Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot recognition and annotation. *Submitted to IEEE TPAMI*, 2014.
8. Y. Fu, Y. Yang, T. Hospedales, T. Xiang, and S. Gong. Transductive multi-label zero-shot learning. In *BMVC*, 2014.
9. Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 2013.
10. C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
11. C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 2013.
12. J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
13. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representation in vector space. In *Proceedings of Workshop at ICLR*, 2013.
14. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.
15. M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *NIPS*, 2013.
16. M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, 2012.
17. M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where–and why semantic relatedness for knowledge transfer. In *CVPR*, 2010.
18. R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. D. Manning, and A. Y. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.
19. X. Wang and Q. Ji. A unified probabilistic approach modeling relationships between attributes and objects. *ICCV*, 2013.
20. F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. *CVPR*, 2013.