Transductive Multi-view Embedding for Zero-Shot Recognition and Annotation

Yanwei Fu, Timothy M. Hospedales, Tao Xiang,
Zhenyong Fu and Shaogang Gong

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

Abstract. Most existing zero-shot learning approaches exploit transfer learning via an intermediate-level semantic representation such as visual attributes or semantic word vectors. Such a semantic representation is shared between an annotated auxiliary dataset and a target dataset with no annotation. A projection from a low-level feature space to the semantic space is learned from the auxiliary dataset and is applied without adaptation to the target dataset. In this paper we identify an inherent limitation with this approach. That is, due to having disjoint and potentially unrelated classes, the projection functions learned from the auxiliary dataset/domain are biased when applied directly to the target dataset/domain. We call this problem the *projection domain shift* problem and propose a novel framework, *transductive multi-view embedding*, to solve it. It is 'transductive' in that unlabelled target data points are explored for projection adaptation, and 'multi-view' in that both low-level feature (view) and multiple semantic representations (views) are embedded to rectify the projection shift. We demonstrate through extensive experiments that our framework (1) rectifies the projection shift between the auxiliary and target domains, (2) exploits the complementarity of multiple semantic representations, (3) achieves state-of-the-art recognition results on image and video benchmark datasets, and (4) enables novel cross-view annotation tasks.

1 Introduction

Humans can distinguish 30,000 basic object classes [3] and many more subordinate ones (e.g. breeds of dogs). To recognise such high number of classes, humans have the ability to "learning to learn" and transfer knowledge from known classes to unknown ones. Inspired by this ability and to minimise the necessary labelled training examples for conventional supervised classifiers, researchers build the recognition models that are capable of classifying novel classes with no training example, i.e. zero-shot learning. The key underpinning idea is to exploit transfer learning via an intermediate-level semantic representation. Specifically, two datasets with disjoint classes are considered: a labelled known auxiliary set where a semantic representation is given for each data point, and a target dataset to be classified with no labelled instance and semantic representation. Such a

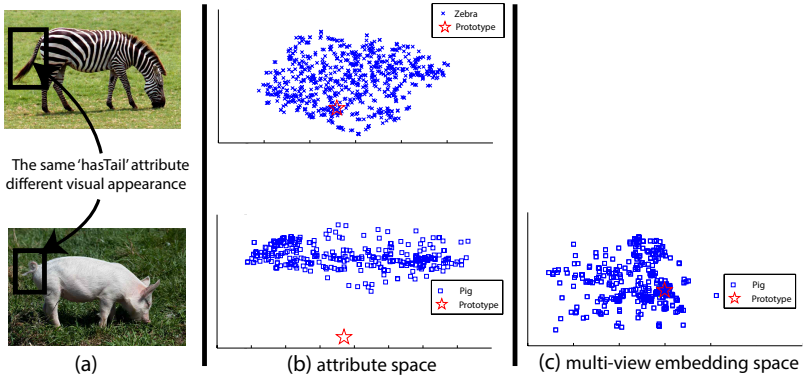

Fig. 1. An example of the projection domain shift problem. Zero-shot prototypes are shown as red stars and image low-level feature projections shown in blue. See text for details.

semantic representation is assumed to be shared between the auxiliary and target datasets. More specifically, apart from class label, each auxiliary data point is labelled by a semantic representation such as visual attributes [18, 6, 21, 11], semantic word vectors [23, 7, 34] or others [28]. A projection function mapping low-level features to the semantic space is learned from the auxiliary dataset by either classification or regression models. Such a projection is then applied directly to map each unlabelled target class instance into the same semantic representation space. Within this semantic space, a zero-shot classifier is pre-defined by "extra-knowledge" to recognise all unseen instances. In particular, a single 'prototype' of each target class is specified in the semantic space. Depending on the semantic space, this prototype can be an attribute annotation vector [18] or a word vector inferred from the target class name [7].

An inherent problem exists in this zero-shot learning approach: Since the two datasets have different and potentially unrelated classes, the underlying semantic prototypes of classes also differ, as 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. We call it the *projection domain shift* problem. This problem is illustrated in Fig. 1, which shows two object classes from the Animals with Attributes (AwA) dataset [20]: Zebra is one of the 40 auxiliary classes whilst Pig is one of 10 target classes. Both of them share the same 'hasTail' attribute, but the visual appearance of their tails differs greatly (Fig. 1(a)). Similarly, many other attributes of Pig are visually very different from those in the 40 auxiliary classes. Fig. 1(b) plots (in 2D using t-SNE [22]) a 85D attribute space representation of the image feature projections and class prototypes (85D binary attribute vectors) to illustrate the existence of the projection domain shift problem: a great discrepancy between the Pig prototype position in the semantic attribute space and the projections

of its class member instances is observed, while the discrepancy does not exist for the auxiliary Zebra class. This discrepancy is caused when the projection functions learned from the 40 auxiliary classes are applied directly to project the Pig instances – what 'hasTail' (as well as the other 84 attributes) visually means is different now. Such a discrepancy will inherently degrade the effectiveness of zero-shot recognition of the Pig class. This projection domain shift problem is uniquely challenging in that there is no labelled information in the target domain to guide domain adaptation in mitigating the problem. To our knowledge, this problem has neither been identified nor addressed in the literature.

In addition to the projection domain shift problem, conventional approaches to zero-shot learning are also limited in exploiting multiple intermediate semantic spaces/views, each of which may contain complementary information. In particular, while both visual attributes [18, 6, 21, 11] and linguistic semantic representations such as word vectors [23, 7, 34] have been independently exploited successfully, it remains unattempted and not straightforward to exploit synergistically multiple semantic 'views'. This is because they are 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 auxiliary data cannot be used directly due to the projection domain shift problem.

In this paper, we propose a transductive multi-view embedding framework to solve both the problems of projection domain shift and synergistic exploitation of multiple semantic views. Specifically, in the first step, each instance of an unlabelled target class 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). To rectify the projection domain shift between auxiliary and target datasets, 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) [13]. The objective of this new embedding space is to transductively (using the unlabelled target data) align each semantic view with each other, and with the low-level feature view to rectify the projection domain shift and exploit their complementarity. This can be seen as an unalignment process and its effects are illustrated by Fig. 1(c), where after alignment, the target Pig class prototype is much closer to its member points in this embedding space, making zero-shot recognition more accurate.

In the second step of our framework, we introduce a novel transductive multi-view Bayesian label propagation (TMV-BLP) algorithm for recognition. This allows us to exploit the manifold of unlabelled test data to compensate for the impoverished supervision available in zero-shot learning, as well as N-shot learning scenario where few target classes instances are available. In particular, a graph is constructed from the projection of each view in the embedding space, plus any available zero-shot 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. Overall our framework has the following

advantages: (1) TMV-BLP can accommodate multiple semantic representations and exploit their complementarity to better rectify the projection domain shift problem and improve recognition. (2) Recognition generalises seamlessly whether none (zero-shot), few (N-shot), ample (fully supervised) examples of the target classes become available. Uniquely it can also synergistically exploit zero + N-shot (i.e., both prototypes and examples) learning. (3) It enables a number of novel cross-view annotation tasks including *zero-shot class description* and *zero attribute learning*. Extensive experiments on benchmark object and video activity datasets demonstrate that our method outperforms state-of-the-art alternatives on both zero-shot and N-shot recognition tasks.

2 Related Work

Semantic spaces for zero-shot learning Learning visual attributes has been topical in the past 5 years. Attribute-centric models have been explored for images [18, 6, 12, 31] and to a lesser extent videos [8, 11, 21]. Most existing studies [18, 17, 25, 26, 30, 37, 1] assume that an exhaustive ontology of attributes has been manually specified at either the class or instance level. However, annotating attributes scales poorly as ontologies tend to be domain specific. This is despite efforts exploring augmented data-driven/latent attributes at the expense of name-ability [6, 21, 11]. To overcome this problem, semantic representations that do not rely on an explicit attribute ontology have been proposed [29, 28], notably *word vectors*. A word space is extracted from linguistic knowledge bases e.g. WordNet or Wikipedia by natural language processing models e.g. [5, 23]. Instead of manually defining an attribute prototype, a novel target class' textual name can be projected into this space and then used as the prototype for zero-shot learning [7, 34]. Importantly, regardless of the semantic space used, existing methods focus on either designing better semantic spaces or how to best learn the projections. The former are orthogonal to our work – any semantic spaces can be used in our framework and better ones would benefit our model. For the latter, no existing work has identified or addressed the projection domain shift problem.

Learning multi-view embedding spaces Relating the low-level feature view and semantic views of data has been exploited in visual recognition and cross modal retrieval. Most existing work [33, 13, 16, 36, 10, 9] focuses on modelling images/videos with associated text (e.g. tags on Flickr/YouTube). Multi-view CCA is often exploited to provide unsupervised fusion of different modalities. However, there are two fundamental differences between previous multi-view embedding work and ours: (1) our embedding space is transductive, that is, learned from unlabelled target data from which all semantic views are estimated by projection rather than being the original views; These projected views thus have the projection domain shift problem that the previous work does not have. (2) The objectives are different: we aim to rectify the projection domain shift via the embedding in order to perform better recognition and annotation while they target primarily cross-modality retrieval.

Multi-view label propagation In most previous zero-shot learning studies (e.g., direct attribute prediction (DAP) [20]), only semantic space prototypes are used for classification. Since the available knowledge (single zero-shot prototype per target class) is very limited, there has been recent interests in additionally exploiting the unlabelled target data by transductive learning [27]. However, apart from suffering from the projection domain shift problem, [27] has limited ability to exploit multiple semantic representations/views. In contrast, after alignment in the embedding space, our framework synergistically integrates the multiple graphs of low-level feature and semantic representations of each instance by transductive multi-view Bayesian label propagation (TMV-BLP). Moreover, TMV-BLP generalises beyond zero-shot to N-shot learning if labelled instances are available for the target classes. Classification on multiple graphs (C-MG) is well-studied in semi-supervised learning. Most solutions are based on the seminal work of Zhou *et al* [38] which generalises spectral clustering from a single graph to multiple graphs by defining a mixture of random walks on multiple graphs. However crucially, the influence/trustworthiness of each graph is given by a weight that has to be pre-defined and its value has a great effect on the performance of C-MG [38]. In this work, we extend the C-MG algorithm in [38] by introducing a Bayesian prior weight for each graph, which can be measured automatically from data. Our experiments show that our TMV-BLP algorithm is superior to [38] and [27].

Our contributions are as follows: (1) To our knowledge, this is the first attempt to investigate and provide a solution to the projection domain shift problem in zero-shot learning. (2) We propose a transductive multi-view embedding space that not only rectifies the projection shift, but also exploits the complementarity of multiple semantic representations of visual data. (3) A novel transductive multi-view Bayesian label propagation algorithm is developed to improve both zero-shot and N-shot learning tasks in the embedding space. (4) The learned embedding space enables a number of novel cross-view annotation tasks.

3 Learning a Transductive Multi-View Embedding Space

Problem setup We have c_S source/auxiliary classes with n_S instances $S = \{X_S, Y_S^i, \mathbf{z}_S\}$ and c_T target classes $T = \left\{X_T, Y_T^i, \mathbf{z}_T\right\}$ with n_T instances. X indicates the $t-$dimensional low-level feature of all instances; so $X_S \subseteq R^{n_s \times t}$ and $X_T \subseteq R^{n_T \times t}$. $\mathbf{z}_S$ and $\mathbf{z}_T$ are the training and test class label vectors. We assume the auxiliary and target classes are disjoint: $\mathbf{z}_S \cap \mathbf{z}_T = \varnothing$. We have I different types of intermediate semantic representations; Y_S^i and Y_T^i represent the $i-th$ type of m_i dimensional semantic representation for the auxiliary and target datasets respectively; so $Y_S^i \subseteq R^{n_S \times m_i}$ and $Y_T^i \subseteq R^{n_T \times m_i}$. Note that for the auxiliary dataset, Y_S^i is given as each data point is labelled. But for the target dataset, Y_T^i is missing, and its prediction $\hat{Y}_T^i$ from X_T is used instead. As we will see later, this is obtained using a projection function learned from the auxiliary dataset. Each target class c has a pre-defined class-level semantic prototype $\mathbf{y}_c^i$ in each semantic view i. In this paper, we consider two types of

intermediate semantic representation (i.e. $I = 2$) – attributes and word vectors, which represent two distinct and complementary sources of information.

We use $\mathcal{X}$, $\mathcal{A}$ and $\mathcal{V}$ to denote the low-level feature, attribute and word vector spaces respectively. The attribute space $\mathcal{A}$ is typically manually defined using a standard ontology. For the word vector space $\mathcal{V}$, we employ the state-of-the-art skip-gram neural network model [23, 24] on all English Wikipedia articles[1] which has higher accuracy and lower computational cost than alternatives such as [5]. Using this learned model, we can project the textual name of any class into the $\mathcal{V}$ space to get its word vector representation. It is a 'free' semantic representation in the sense that the generating process does not need any human annotations. We next address how to project low-level features into these spaces.

Learning the projections of semantic spaces Mapping images and videos into a semantic space i requires a projection function $f^i : \mathcal{X} \rightarrow \mathcal{Y}^i$. This is typically realised by classifiers [18] or regressors [34]. In this paper, using the auxiliary set S, we train support vector classifiers $f^{\mathcal{A}}(\cdot)$ and support vector regressors $f^{\mathcal{V}}(\cdot)$ for each dimension of the attribute and word vectors respectively. Then the target class instances X_T have the semantic projections: $\hat{Y}_T^{\mathcal{A}} = f^{\mathcal{A}}(X_T)$ and $\hat{Y}_T^{\mathcal{V}} = f^{\mathcal{V}}(X_T)$. However, these predicted intermediate semantics have the projection domain shift problem illustrated in Fig. 1. To solve this, we learn a transductive multi-view semantic embedding space to align the semantic projections with the low-level features of target data.

Learning transductive multi-view semantic embedding To learn an embedding space capable of rectifying the domain shift, we employ multi-view Canonical Correlation Analysis (CCA) for E views, each denoted as $\boldsymbol{\Phi}^i$. Specifically, in this work we project three views of each target class instance $f^{\mathcal{A}}(X_T)$, $f^{\mathcal{V}}(X_T)$ and X_T (i.e. $E = I + 1 = 3$) into a shared embedding space and the three projection functions W^i are learned by

$$\begin{aligned} \min \quad & \sum_{i,j=1}^{E} \quad Trace(W^i \Sigma_{ij} W^j) \\ = \quad & \sum_{i,j=1}^{E} \quad \| \boldsymbol{\Phi}^i W^i - \boldsymbol{\Phi}^j W^j \|_F^2 \\ \text{s.t.} \ & \left[W^i\right]^T \Sigma_{ii} W^i = I \ \left[\mathbf{w}_k^i\right]^T \Sigma_{ij} \mathbf{w}_l^j = 0 \\ i \neq j, k \neq l \quad & i, j = 1, \cdots, E \quad k, l = 1, \cdots, n_T \end{aligned} \tag{1}$$

where W^i is the projection matrix which maps the view $\boldsymbol{\Phi}^i$ (a n_T row matrix) into the embedding space and $\mathbf{w}_k^i$ is the kth column of W^i. Σ_{ij} is the covariance matrix between $\boldsymbol{\Phi}^i$ and $\boldsymbol{\Phi}^j$. The dimensionality of the embedding space is the sum of that of $\boldsymbol{\Phi}^i$ – there is obviously feature redundancy. Since the importance of each dimension is reflected by its corresponding eigenvalue [14, 13, 4], we use the eigenvalues to weight the dimensions and define a *weighted embedding space* Γ:

$$\boldsymbol{\Psi}^i = \boldsymbol{\Phi}^i W^i \left[D^i\right]^{\lambda} = \boldsymbol{\Phi}^i W^i \tilde{D}^i, \tag{2}$$

[1] Only articles are used without any user talk/discussion. To 13 Feb. 2014, it includes 2.9 billion words and 4.33 million vocabulary (single and bi/tri-gram words). It is downloadable from Yanwei's website.

where D^i is a diagonal matrix with its diagonal elements set to the eigenvalues of each dimension in the embedding space, λ is a power weight of D^i and empirically set to 4 [13], and $\boldsymbol{\Psi}^i$ is the final representation of data from view i in Γ. In this work, three views are considered; for notational convenience, we index $i \in \{\mathcal{X}, \mathcal{V}, \mathcal{A}\}$. The same formulation can be used if more than three views are available.

Similarity in the embedding space The choice of similarity metric is important for the high-dimensional embedding spaces Γ [13]. In particular, extensive evidences in text analysis and information retrieval have shown that high-dimensional embedding vectors are naturally directional and using cosine similarity provides significant robustness against noise [2, 13, 14]. Therefore for the subsequent recognition and annotation tasks, we compute cosine similarity in Γ by l_2 normalisation: normalising any vector $\psi_k^i \in \boldsymbol{\Psi}^i$ to unit length (i.e. $\| \psi_k^i \|_2 = 1$). Thus cosine similarity is given by the inner product of any two vectors in Γ. Finally, equipped with a weighted and normalised embedding space Γ, any two vectors can be directly compared no matter whether the original view is $\mathcal{X}$, $\mathcal{A}$ or $\mathcal{V}$.

4 Recognition by Multi-view Bayesian Label Propagation

We now introduce a unified framework for exploiting unlabelled target data transductively to improve zero-shot recognition as well as N-shot learning if sparse examples are available. We assume a target class c has a prototype $\mathbf{y}_c^i$ (either a manual binary attribute vector, or the class name representation in the word space as a word vector) in each semantic view for zero-shot, and/or a few labelled instances for N-shot classification. To exploit the learned embedding Γ for recognition, we project three views of each target class instance $f^{\mathcal{A}}(X_T)$, $f^{\mathcal{V}}(X_T)$ and X_T as well as the target class prototypes into this space. The prototypes $\mathbf{y}_c^i$ for views $i \in \{\mathcal{A}, \mathcal{V}\}$ are projected as $\psi_c^i = \mathbf{y}_c^i W^i \tilde{D}^i$. So we have $\psi_c^{\mathcal{A}}$ and $\psi_c^{\mathcal{V}}$ for the attribute and word vector prototypes of each target class c in Γ. In the absence of a prototype for the (non-semantic) low-level feature view $\mathcal{X}$, we synthesise it as $\psi_c^{\mathcal{X}} = (\psi_c^{\mathcal{A}} + \psi_c^{\mathcal{V}})/2$.

Most if not all of the target class instances are unlabelled. To exploit the unlabelled data transductively for classification, we consider graph-based semi-supervised learning in Γ. However, since our embedding space contains multiple projections of the target data, it is hard to define a single graph that exploits the manifold structure of each view. We therefore consider the graphs defined by the projection of each view in a multi-view label propagation algorithm (TMV-BLP). Thanks to the shared embedding space Γ, these heterogeneous graphs become comparable and can be connected by a Bayesian prior weight estimated from data. TMV-BLP provides multi-view label propagation by unifying the multi-graph into a single graph via a random walk within and across the graphs. The initial label information from the prototypes (zero-shot) and/or the few labelled target data points (N-shot learning) is then propagated to the unlabelled data.

Multi-view Bayesian graph construction In Γ we aim to build a graph $\mathcal{G}$ relating labelled and unlabelled data and prototypes. Each view projection defines a node, and the distance between any pair of nodes is:

$$\omega(\psi_k^i, \psi_l^j) = \exp(\frac{< \psi_k^i, \psi_l^j >^2}{\delta}) \tag{3}$$

where $< \psi_k^i, \psi_l^j >^2$ is the square of inner product between the $i-th$ and $j-th$ projections of nodes k and l with a free parameter[2] δ. However, exhaustively connecting all projections of all data is computationally expensive. To balance efficiency and reflecting the topological manifold structure of the graphs, we simplify Eq (3) by two strategies: (1) We construct a k-nearest-neighbour graph $\mathcal{G}^i$ within each projection i, i.e., $i = j$ and $k \neq l$ in Eq (3) with $K = 30$ nearest neighbours[3]. (2) To connect heterogeneous graphs $\mathcal{G}^i$ and $\mathcal{G}^j$ $(i \neq j)$, we only compute the similarity across projections at the same data point ($k = l$ but $i \neq j$ in Eq (3)).

To propagate label information from labelled nodes to other target instances, a classic strategy is random walks [38]. We define a random walk process within and across graphs. A natural random walk within $\mathcal{G}^i$ for two nodes k and l has the following transition probability,

$$p(k \to l|\mathcal{G}^i) = \frac{\omega(\psi_k^i, \psi_l^i)}{\sum_m \omega(\psi_k^i, \psi_m^i)}, \tag{4}$$

and the stationary probability for node k,

$$\pi(k|\mathcal{G}^i) = \frac{\sum_l \omega(\psi_k^i, \psi_l^i)}{\sum_k \sum_l \omega(\psi_k^i, \psi_l^i)}. \tag{5}$$

To connect the graphs across views $i \neq j$, we need to model the overall graph probability. Let $p(\mathcal{G}^i)$ denote the prior probability of graph $\mathcal{G}^i$ in the random walk. This prior reflects how informative $\mathcal{G}^i$ is. Then the posterior probability to choose graph $\mathcal{G}^i$ at projection/node ψ_k^i will be:

$$p(\mathcal{G}^i|k) = \frac{\pi(k|\mathcal{G}^i)p(\mathcal{G}^i)}{\sum_i \pi(k|\mathcal{G}^i)p(\mathcal{G}^i)}. \tag{6}$$

For any pair of nodes k and l, the transition probability across multiple graphs can be computed by Bayesian model averaging,

$$p(k \to l) = \sum_i p(k \to l|\mathcal{G}^i) \cdot p(\mathcal{G}^i|k). \tag{7}$$

[2] Most previous work [27, 38] needs to do cross validation for δ. Inspired by [19], a rule of thumb for setting δ is $\delta \approx \underset{k,l=1,\cdots,n}{\text{median}} < \psi_k^i, \psi_l^j >^2$ to balance roughly the same number of similar as dissimilar example pairs. This makes the edge weight invariant to the value scale of the heterogeneous graph.

[3] It can be varied from $10 \sim 50$ with little effects in our experiments.

In addition, the stationary probability across multiple graphs is computed as:

$$\pi(k) = \sum_i \pi(k|\mathcal{G}^i) \cdot p(\mathcal{G}^i). \tag{8}$$

Finally, the prior probability of each graph $p(\mathcal{G}^i)$ is computed as

$$p(\mathcal{G}^i) = \frac{\sum_k \sum_{j\in\{\mathcal{X},\mathcal{V},\mathcal{A}\},j\neq i} \omega(\psi_k^i, \psi_k^j)}{\sum_k \sum_{r\in\{\mathcal{X},\mathcal{V},\mathcal{A}\}} \sum_{u\in\{\mathcal{X},\mathcal{V},\mathcal{A}\},r\neq u} \omega(\psi_k^r, \psi_k^u)} \tag{9}$$

The intuition is that this is the relative consensus of each graph to all the others. A graph that is on average similar/in consensus with the others gets a stronger prior compared to an outlying graph which is not in consensus with others.

Label propagation Given the graph and random walk process defined above, we can derive the semi-supervised label propagation. Let P denote the transition probability matrix defined by Eq (7) and Π the diagonal matrix with the elements $\pi(k)$ computed by Eq (8). The Laplacian matrix $\mathcal{L}$ combines information of different views and is defined as: $\mathcal{L} = \Pi - \frac{\Pi P + P^T \Pi}{2}$. The label matrix Z for labelled N-shot data or zero-shot prototypes is defined as:

$$Z(q_k, c) = \begin{cases} 1 & q_k \in class\, c \\ -1 & q_k \notin class\, c \\ 0 & unknown \end{cases} \tag{10}$$

Given the label matrix Z and Laplacian $\mathcal{L}$, label propagation on multiple graphs has the closed-form solution: $\hat{Z} = \eta(\eta\Pi + \mathcal{L})^{-1}\Pi Z$ where η is a regularisation parameter[4]. Note that in our framework, both labelled target class instances and prototypes are modelled as graph nodes. Thus the difference between zero-shot and N-shot learning lies only on the initial labelled instances: Zero-shot learning has the prototypes as labelled nodes; N-shot has instances as labelled nodes; and a new condition exploiting both prototypes and N-shot together is possible. This unified recognition framework thus applies when either or both of prototypes and labelled instances are available.

5 Annotation and Beyond

Our multi-view embedding space Γ bridges the semantic gap between low-level features $\mathcal{X}$ and semantic representations $\mathcal{A}$ and $\mathcal{V}$. Leveraging this cross-view mapping, annotation [15, 36, 13] can be improved and applied in novel ways. We consider three annotation tasks here.

Instance level annotation via attribute classification Given a new instance u, we can describe/annotate it by predicting its attributes. The conventional solution is directly applying $\hat{\mathbf{y}}_u^{\mathcal{A}} = f^{\mathcal{A}}(\mathbf{x}_u)$ for test data. However as we have shown this suffers from projection domain shift problem. With the

[4] It can be varied from $1 - 10$ with little effects in our experiments

learned embedding Γ, we can now infer attributes for each target class instance $\hat{\mathbf{y}}_u^{\mathcal{A}} = \mathbf{x}_u W^{\mathcal{X}} \tilde{D}^{\mathcal{X}} \left[W^{\mathcal{A}} \tilde{D}^{\mathcal{A}} \right]^{-1}$.

Zero-shot class description From a broader pattern recognition perspective, one might be interested to ask what are the attributes of an unknown class, based solely on the name of the class. This *zero-shot class description* task could be useful, for example, to hypothesise the zero-shot attribute prototype of a class instead of defining it by experts [18] or ontology [11]. Our transductive embedding space enables this task by connecting semantic word space (i.e. naming) and discriminative attribute space (i.e. describing). Therefore, given the prototype $\mathbf{y}_c^{\mathcal{V}}$ from the name of a novel target class c, we compute $\hat{\mathbf{y}}_c^{\mathcal{A}} = \mathbf{y}_c^{\mathcal{V}} W^{\mathcal{V}} \tilde{D}^{\mathcal{V}} \left[W^{\mathcal{A}} \tilde{D}^{\mathcal{A}} \right]^{-1}$ to generate their class-level attribute description.

Zero attribute learning This task is the inverse of the previous task: *inferring the name of a class given a set of attributes*. It is useful, for example, to validate or assess a proposed zero-shot attribute prototype, or to provide an automated semantic-property based index into a dictionary or database. To our knowledge, this is the first attempt for evaluating the quality of a class attribute prototype because no previous work has directly and systematically linked linguistic knowledge space with visual attribute space. Specifically given an attribute prototype $\mathbf{y}_c^{\mathcal{A}}$, we can use $\hat{\mathbf{y}}_c^{\mathcal{V}} = \hat{\mathbf{y}}_c^{\mathcal{A}} W^{\mathcal{A}} \tilde{D}^{\mathcal{A}} \left[W^{\mathcal{V}} \tilde{D}^{\mathcal{V}} \right]^{-1}$ to name the corresponding class and do retrieval on dictionary words in $\mathcal{V}$ using $\hat{\mathbf{y}}_c^{\mathcal{V}}$.

6 Experiments

Datasets and settings We evaluate our framework on two widely used image/video attribute datasets: Animal with Attribute (AwA) and Unstructured Social Activity Attribute (USAA). AwA [18] provides 50 classes of animals (30475 images) and 85 associated class-level attributes (such as furry, and hasClaws). It provides a defined source/target split for zero-shot learning with 10 classes and 6180 images held out. USAA is a video dataset [8, 11] with 69 instance-level attributes for 8 classes of complex social group activity videos from YouTube. Each class has around 100 training and testing videos respectively. USAA provides instance-level attributes since there are significant intra-class variabilities. We use the thresholded mean of instances from each class to define a binary attribute prototype as in [11]. We use the same transfer learning setting in [11]: 4 classes as source and 4 classes as target data. We used exactly the same RGB colour histograms, SIFT, rgSIFT, PHOG, SURF and local self-similarity histograms in [18] for AwA, and SIFT, MFCC and STIP as low-level features for USAA as in [8]. We report absolute classification accuracy on USAA and mean accuracy for AwA for direct comparison to published results. The word vector space is trained by the skip model [23] with 100 dimensions[5].

[5] All codes/features are downloadable from Yanwei's website: `http://www.eecs.qmul.ac.uk/~yf300/embedding/`.

Approach	AwA	USAA
TMV-BLP	**47.1**	**47.8**
DAP	40.5([18]) / 41.4([20]) / 38.4*	33.2([11]) / 35.2*
IAP	27.8([18]) / 42.2([20])	–
M2LATM [11]	41.3	41.9
ALE/HLE/AHLE [1]	37.4/39.0/43.5	–
Mo/Ma/O/D [29]	27.0 / 23.6 / 33.0 / 35.7	–
PST [27]	42.7	36.2*
[35]	43.4	–

Table 1. Comparison with the state-of-the-art on zero-shot learning on AwA and USAA. 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 [29] respectively. Note *: our implementation.

6.1 Evaluating Zero-Shot Learning

Comparisons with state-of-the-art Our method (TMV-BLP) is compared against the state-of-the-art models for zero-shot learning in Table 1. For fair comparison, human effort exploited by all compared methods is limited to attribute annotation as in [20, 8]. This excludes methods such as [37] which require additional human interventions. Note that our semantic vectors are 'zero' cost for human annotations, because they are generated by projecting classes' textual name into the $\mathcal{V}$ space. Apart from our method, the AHLE method in [1] also use two semantic spaces: attribute and WordNet hierarchy. Different from our embedding framework, AHLE simply concatenates the two spaces. Our TMV-BLP outperforms all the other methods by a noticeable margin on both datasets, showing the effectiveness of our approach.

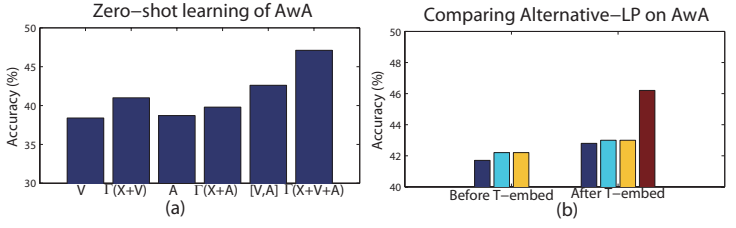

Fig. 2. (a)Effectiveness of transductive multi-view embedding for zero-shot learning on AwA and USAA. $[\mathcal{V}, \mathcal{A}]$ indicates the concatenation of semantic word and attribute space vectors. $\Gamma(\mathcal{X}+\mathcal{V})$ and $\Gamma(\mathcal{X}+\mathcal{A})$ mean using low-level+semantic word spaces and low-level+attribute spaces respectively to learn the embedding. $\Gamma(\mathcal{X}+\mathcal{V}+\mathcal{A})$ indicates using all 3 views to learn the embedding.(b) Comparing alternative label propagation methods. Note: T-embed means Transductive embedding spaces.

Effectiveness of transductive multi-view embedding We validate the contribution of our transductive multi-view embedding space by splitting up and comparing the results of different combinations in Fig. 2 (a): $\mathcal{V}$ vs. $\Gamma(\mathcal{X}+\mathcal{V})$,

$\mathcal{A}$ vs. $\Gamma(\mathcal{X}+\mathcal{A})$ and $[\mathcal{V},\mathcal{A}]$ vs. $\Gamma(\mathcal{X}+\mathcal{V}+\mathcal{A})$ (see the caption of Fig. 2(a) for definitions). We use DAP for $\mathcal{A}$ and nearest neighbour for $\mathcal{V}$ and $[\mathcal{V},\mathcal{A}]$, because the prototypes of $\mathcal{V}$ are not binary vectors so DAP cannot be applied. We use TMV-BLP for $\Gamma(\mathcal{X}+\mathcal{V})$ and $\Gamma(\mathcal{X}+\mathcal{A})$. We highlight the following observations: (1) After transductive embedding, $\Gamma(\mathcal{X}+\mathcal{V}+\mathcal{A})$, $\Gamma(\mathcal{X}+\mathcal{V})$ and $\Gamma(\mathcal{X}+\mathcal{A})$ outperform $[\mathcal{V},\mathcal{A}]$, $\mathcal{V}$ and $\mathcal{A}$ respectively. This means that the transductive embedding is helpful whichever semantic space is used; and validates the effectiveness of the embedding space in rectifying the projection domain shift by aligning the semantic views with low-level features. (2) The results of $[\mathcal{V},\mathcal{A}]$ are higher than those of $\mathcal{A}$ and $\mathcal{V}$ individually, showing that the two semantic views are indeed complementary. However, our TMV-BLP on all views $\Gamma(\mathcal{X}+\mathcal{V}+\mathcal{A})$ further improves individual embeddings $\Gamma(\mathcal{X}+\mathcal{V})$ and $\Gamma(\mathcal{X}+\mathcal{A})$.

Comparison with alternative label propagation methods We also compare two alternative label propagation methods: C-MG [38] and PST [27]. We use equal weights for each graph for C-MG and the same parameters from [27] for PST. We compare all methods before and after the embedding, as shown in Fig. 2(b). The performance of [27] depends on good quality initial labels while our TMV-BLP uses the trivial initial labels in Eq (10). We conclude that: (1) TMV-BLP in our embedding space outperforms both alternatives. (2) The embedding also improves C-MG and PST, due to alleviated projection domain shift via aligning the semantic projections and low-level features. This result shows that both the proposed embedding space and the Bayesian label propagation algorithm contributes to the superior performance of our method.

6.2 Evaluating N-Shot Learning

We next consider N-shot learning on AwA and USAA varying the number of training target class instances. This is challenging when there are few training examples per target class. We also consider the situation [27] where both a few training examples *and* a zero-shot prototype may be available (denoted with suffix +), and contrast it to conventional N-shot learning (denoted with suffix −). Note that our TMV-BLP can be used in both conditions but the PST method [27] mainly applies to the + condition[6]. All experiments use the same training instances and are repeated for 10 rounds to reduce variance. Evaluation is done on the remaining available images from the test split after removing the N instances. From the results shown in Fig. 3, we have the following observations and conclusions: (1) TMV-BLP+ always achieves the best performance, particularly given few training examples. This shows the effectiveness of our framework by combining complementary semantic and low-level feature information. We note that with 50 labelled instances per target class (fully supervised), using SVM with RBF kernel in the embedded space Γ obtains the same results as our TMV-BLP+, because the transductive kernel matrix (inverse of the Laplacian matrix $\mathcal{L}$) essentially models the same information as the SVM kernel matrix [32]. (2) As clearly shown in the AwA results, PST+ outperforms TMV-BLP-

[6] PST- corresponds to the standard label propagation (see Fig 3(b) in [27]).

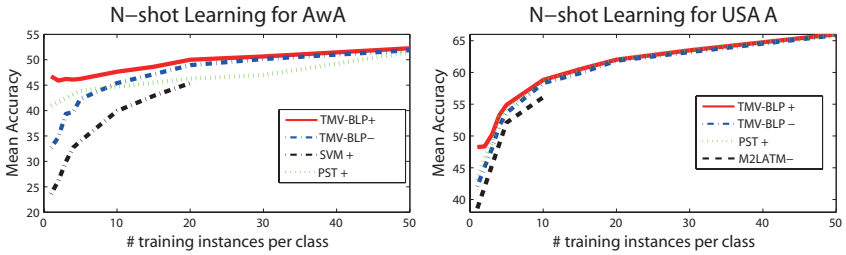

Fig. 3. N-shot learning comparison. PST+ is the method in [27] which uses prototypes for the initial label matrix. SVM+ and M2LATM- are the SVM and M2LATM methods used in [20] and [11] respectively. For fair comparison, we modify the SVM- used in [20] into SVM+.

AwA	raccoon	giant panda	humpback+whale	rat
top-5	**furry**, **quadrupedal**, **tail**, **nest spot**, **tree**	**vegetation**, **furry**, hops, **grazer**, **quadrupedal**.	nest spot, **slow**, **tail**, quadrupedal, weak	**tail**, nest spot, weak, **quadrupedal**, grazer
bot-5	*smart*, *bipedal*, *swims*, *tough skin*, *hairless*	*swims*, *ocean*, *hairless*, new world	*claws*, *hands*, big, new world, *bipedal*	*tusks*, small, new world, *bipedal*

Table 2. Ranking attributes of AwA unseen testing classes. Bold font illustrates true positives; italic illustrates true negatives.

with less than 10 instances per class because PST+ exploits the prototypes. This suggests that a single good prototype is more informative than a few labelled instances in N-shot learning. This also explains when only few training labels are observed why the N-shot learning results of TMV-BLP+ are worse than its zero-shot learning results. (3) Nevertheless, TMV-BLP- still surpasses PST+ with more training instances because TMV-BLP combines the different views of the training instances, and the strong effect of the prototype is outweighed as more labelled instances become available.

6.3 Evaluating Annotation

Instance annotation by attributes To quantify the annotation performance, we predict attributes/annotations for each target class instance for USAA. We employ two standard measures: mean average precision (mAP) and F-measure (FM) between the estimated and true annotation list. Using our three-view embedding space, our method (FM:0.341, mAP: 0.355) outperforms significantly the baseline of directly estimating $\mathbf{y}_u^{\mathcal{A}} = f^{\mathcal{A}}(\mathbf{x}_u)$ (FM:0.299, mAP: 0.267).

Novel annotation tasks beyond visual recognition We next illustrate two novel annotation tasks. In the *Zero-Shot Description* task, we explicitly infer the member attributes, given only the textual name of a novel class. Table 2 illustrates this for AwA by showing that the top/bottom 5 attributes associated with each class are meaningful (in ideal cases, all top 5 should be true positives and all bottom 5 true negatives). In the *Zero-Attribute Learning* task we attempt

the reverse, inferring class names given a list of attributes. Table 3(a) shows the query attributes used for USAA (note that class name is shown for brevity, but it is the attributes of the classes that are queried) and the top-4 ranked list of classes returned. We emphasise the average rank of the true class is an impressive 2.13 (out of 4.33M vocabulary with chance-level 2.3×10^{-7}), compared with the average rank of 110.24 by directly querying word space [23] by using the textual descriptions of attributes. Table 3(b) shows an example of "incremental" query of using ontology definition of birthday party [11]. We firstly query by the *wrapped presents* attribute only, followed by adding *small balloon* and all the other attributes (*birthday songs* and *birthday caps*). The changing list of top ranked retrieved words intuitively reflects the expectation of the combinatorial meaning of the attributes.

(a)	Query via embedding space	Query attribute words in word space
g	**party**, **graduation**, audience, caucus	cheering, proudly, dressed, wearing
m	**music**, **performance**, musical, heavy metal	sing, singer, sang, dancing
w_c	**w_c**, wedding, glosses, stag	nun, christening, bridegroom, **w_c**

(b) Attribute Query	Top Ranked Words
wrapped presents	music; performance; solo_performances;performing
+small balloon	wedding; wedding_reception; birthday_celebration; birthday
+All attibutes	**birthday_party**; prom; wedding reception

Table 3. (a) Querying by attributes of classes. g,m and w_c indicate graduation party, music_performance and wedding_ceremony respectively. (b) An incrementally constructed query for birthday party. Bold indicates true positive words retrieved.

7 Conclusions and future work

We identified the challenge of projection domain shift in zero-shot learning and presented a new framework TMV-BLP to solve it by rectifying the biased projections in an embedding space. TMV-BLP synergistically exploits multiple intermediate semantic representations, as well as the manifold structure of unlabelled test data to improve recognition in a unified way for both zero and N-shot learning tasks. So we achieve state-of-the-art performance on the challenging AwA and USAA datasets. Finally, we demonstrate that our framework enables novel tasks of relating textual class names and their semantic attributes.

References

1. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for attribute-based classification. In: CVPR (2013)
2. Banerjee, A., Dhillon, I.S., Ghosh, J., Sra, S.: Clustering on the unit hypersphere using von mises-fisher distributions. JMLR (2005)

3. Biederman, I.: Recognition by components - a theory of human image understanding. Psychological Review (1987)
4. Blitzer, J., Foster, D.P., Kakade, S.M.: Zero-shot domain adaptation: A multi-view approach (2009)
5. Brown, P.F., Pietra, V.J., V.deSouza, P., C.Lai, J., L.Mercer, R.: Class-based n-gram models of natural language. Journal Computational Linguistics (1992)
6. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR (2009)
7. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model andrea. In: NIPS (2013)
8. Fu, Y., Hospedales, T., Xiang, T., Gong, S.: Attribute learning for understanding unstructured social activity. In: ECCV (2012)
9. Fu, Y.: Multi-view metric learning for multi-view video summarization. http://arxiv.org/abs/1405.6434 (2014)
10. Fu, Y., Guo, Y., Zhu, Y., Liu, F., Song, C., Zhou, Z.H.: Multi-view video summarization. IEEE TMM 12(7), 717–729 (2010)
11. Fu, Y., Hospedales, T.M., Xiang, T., Gong, S.: Learning multi-modal latent attributes. TPAMI (2013)
12. Fu, Y., Hospedales, T.M., Xiang, T., Yao, Y., Gong, S.: Interestingness prediction by robust learning to rank. In: ECCV (2014)
13. Gong, Y., Ke, Q., Isard, M., Lazebnik, S.: A multi-view embedding space for modeling internet images, tags, and their semantics. IJCV (2013)
14. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis; an overview with application to learning methods. In: Neural Computation (2004)
15. Hospedales, T., Gong, S., Xiang, T.: Learning tags from unsegmented videos of multiple human actions. In: ICDM (2011)
16. Hwang, S.J., Grauman, K.: Learning the relative importance of objects from tagged images for retrieval and cross-modal search. IJCV (2011)
17. Hwang, S.J., Sha, F., Grauman, K.: Sharing features between objects and their attributes. In: CVPR (2011)
18. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
19. Lampert, C.H.: Kernel methods in computer vision. Foundations and Trends in Computer Graphics and Vision (2009)
20. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE TPAMI (2013)
21. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: CVPR (2011)
22. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-sne. JMLR (2008)
23. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representation in vector space. In: Proceedings of Workshop at ICLR (2013)
24. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., , Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS (2013)
25. Palatucci, M., Hinton, G., Pomerleau, D., Mitchell, T.M.: Zero-shot learning with semantic output codes. In: NIPS (2009)
26. Parikh, D., Grauman, K.: Relative attributes. In: ICCV (2011)
27. Rohrbach, M., Ebert, S., Schiele, B.: Transfer learning in a transductive setting. In: NIPS (2013)

28. Rohrbach, M., Stark, M., Schiele, B.: Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In: CVPR (2012)
29. Rohrbach, M., Stark, M., Szarvas, G., Gurevych, I., Schiele, B.: What helps where– and why semantic relatedness for knowledge transfer. In: CVPR (2010)
30. Scheirer, W.J., Kumar, N., Belhumeur, P.N., Boult, T.E.: Multi-attribute spaces: Calibration for attribute fusion and similarity search. In: CVPR (2012)
31. Shi, Z., Hospedales, T.M., Xiang, T.: Weakly supervised object-attribute prediction and localisation. In: ECCV (2014)
32. Smola, A.J., Kondor, R.: Kernels and regularization on graphs. In: Proc. 16th Annual Conference on Learning Theory (2003)
33. Socher, R., Fei-Fei, L.: Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In: CVPR (2010)
34. Socher, R., Ganjoo, M., Sridhar, H., Bastani, O., Manning, C.D., Ng, A.Y.: Zero-shot learning through cross-modal transfer. In: NIPS (2013)
35. Wang, X., Ji, Q.: A unified probabilistic approach modeling relationships between attributes and objects. ICCV (2013)
36. Wang, Y., Gong, S.: Translating topics to words for image annotation. In: ACM CIKM (2007)
37. Yu, F.X., Cao, L., Feris, R.S., Smith, J.R., Chang, S.F.: Designing category-level attributes for discriminative visual recognition. CVPR (2013)
38. Zhou, D., Burges, C.J.C.: Spectral clustering and transductive learning with multiple views. ICML 07 (2007)