This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2737007, IEEE Transactions on Pattern Analysis and Machine Intelligence

Zero-Shot Learning on Semantic Class Prototype Graph

Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong

Abstract—Zero-Shot Learning (ZSL) for visual recognition is typically achieved by exploiting a semantic embedding space. In such a space, both seen and unseen class labels as well as image features can be embedded so that the similarity among them can be measured directly. In this work, we consider that the key to effective ZSL is to compute an optimal distance metric in the semantic embedding space. Existing ZSL works employ either Euclidean or cosine distances. However, in a high-dimensional space where the projected class labels (prototypes) are sparse, these distances are suboptimal, resulting in a number of problems including hubness and domain shift. To overcome these problems, a novel manifold distance computed on a semantic class prototype graph is proposed which takes into account the rich intrinsic semantic structure, i.e., semantic manifold, of the class prototype distribution. To further alleviate the domain shift problem, a new regularisation term is introduced into a ranking loss based embedding model. Specifically, the ranking loss objective is regularised by unseen class prototypes to prevent the projected object features from being biased towards the seen prototypes. Extensive experiments on four benchmarks show that our method significantly outperforms the state-of-the-art.

Index Terms—Zero-shot learning, semantic embedding, class prototype graph, hubness, semantic manifold, absorbing Markov chain process

1 Introduction

A recent trend in visual recognition research is to scale up the number of object categories. However, most existing recognition models are based on supervised learning and require a large number (at least 100s) of training samples to be collected and annotated for each object class to capture its intra-class appearance variations. This severely limits their scalability – collecting images of common objects such as chairs is easy, but many other categories are rare, e.g. a newly identified specie of beetles on a remote pacific island. None of these models can work with few or even no training samples for a given class. This is one of the reasons why the popular large-scale visual recognition challenge (ILSVRC) [54] mainly focuses on the task of recognising 1K classes, a rather small subset of the ImageNet dataset of which there are in total 21,814 classes with 14M images. The difficulty is that many object classes of the larger ImageNet dataset are only composed of a handful of images including 296 classes with only one image. In this wider context, scalability poses a critical challenge to large-scale visual recognition.

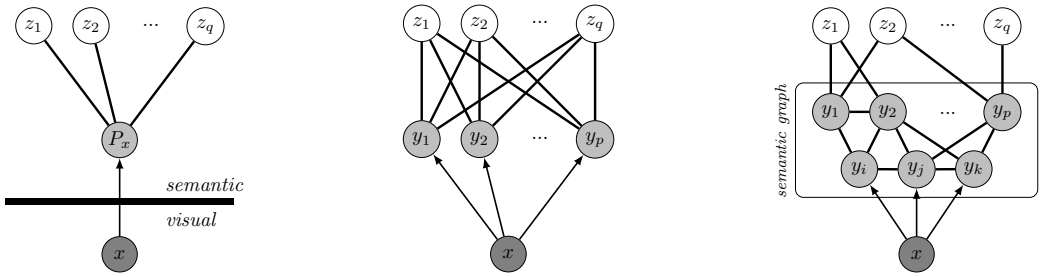
Humans can identify approximately 30,000 basic object categories [7] and many more sub-classes, e.g. breeds of dogs and combination of attributes and objects. Importantly, humans are very good at recognising objects without seeing any visual samples. In machine learning, this is considered as the problem of *zero-shot learning* (ZSL). For example, a child would have no problem recognising a zebra if he/she has seen horses before and read somewhere that a zebra is but a horse with black-and-white stripes. Inspired by humans' ZSL ability, recently there is a surge of interest in machine ZSL for scaling up visual recognition [20], [34], [46], [51], [44], [3], [52], [29], [31], [26], [36], [66], [64], [9].

The reason why humans can perform ZSL is because there exist language knowledge bases, e.g. books, Wikipedia, which provide high-level/semantic description of a new/unseen class (zebra) and make connection between it and seen classes and visual concepts (horse, stripe). Similarly machine zero-shot recognition relies on the existence of a labelled training set of seen classes and the knowledge about how each unseen class is semantically related to the seen classes. Seen and unseen classes are usually related in a high dimensional vector space, which is called semantic embedding space. Such a space can be a semantic attribute space [34], [19] or a semantic word vector space [22], [44], [59]. In the semantic embedding space, the names of both seen and unseen classes are embedded as vectors called class prototypes [23]. The semantic relationships between classes can then be measured by a distance, e.g. the prototypes of zebra and horse should be close to each other. Importantly, the same space can be used to embed a feature representation of an object image, making visual recognition possible.

Specifically, almost all existing ZSL methods adopt a **Semantic Embedding** (**SE**) approach (Fig. 1(a)). First, a projection function between the visual feature space and the semantic embedding space is learned using the labelled training visual data consisting of seen classes only. This function is then used to project/embed the visual repre-

• *Z.Fu, T. Xiang, E. Kodirov and S. Gong are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, United Kingdom.*
E-mail: {z.fu, t.xiang, e.kodirov, s.gong}@qmul.ac.uk

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2737007, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) SE (b) SR (c) SR (SE) + AMP

Fig. 1: A semantic manifold distance model unifies Semantic Embedding (SE) and Semantic Relatedness (SR) based methods for ZSL. Given an unseen class image, x and P_x are the visual feature vector and its projection in the embedding space respectively. The seen and unseen class prototypes are denoted as y and z respectively.

sentation of each unseen class image into the space where the unseen class prototypes also reside. The final step of recognition is typically based on simple nearest neighbour (NN) – the class is determined by the nearest unseen class prototype[1]. Although rarely used recently, there exists a second approach called **Semantic Relatedness** (**SR**) [34] (Fig. 1(b)). Taking this approach, a n-way discrete classifier for the seen classes is first learned, which is then used to compute a visual similarity vector between an image of unseen class and those of the seen classes [4], [51]. The semantic relatedness between the seen and unseen classes is measured by the distance between their prototypes. The resultant semantic relatedness (similarity) vector is then compared with the visual similarity vector and the image is classified to an unseen class if the two types of similarities match as the closest by NN.

For either SE or SR, measuring the similarity between prototype vectors in the semantic embedding space for NN search is the key. However, most existing works on ZSL focus on learning the best semantic embedding space or the projection function from the feature to the embedding space. When it comes to computing distance/measuring similarity in the embedding space, they simply use Euclidean or cosine distance. This results in two major problems: (1) **Hubness** – in a high dimensional space, nearest neighbour suffers from the existence of hubs, i.e. the class prototypes which are the nearest neighbours of many test data points, regardless which classes they belong to [47]. The problem is intrinsic to a high-dimensional vector space when NN search is performed. Although the semantic space used in ZSL may not have a particularly high dimension, the number of unseen class prototypes is normally small therefore aggravating the hubness problem. (2) **Domain shift** – for a SE-based approach, this is the projection domain shift problem [24]; that is, since the projection for visual feature embedding is learned from the seen classes, the projected unseen class data points would be biased towards the seen classes. As a result, they could be far away from their corresponding unseen class prototypes, making hubs easier to emerge and directly measuring similarity using Euclidean/cosine distance less meaningful. Although this bias does not occur for a SR-based approach, by which no explicit feature embedding is necessary, another form of domain shift, the visual-semantic domain shift takes its place – visually similar objects may not be semantically similar, e.g., an orange and a tennis ball are visually similar but semantically distinct. A NN search based on a simple distance such as Euclidean or cosine would thus suffer from both types of domain shift.

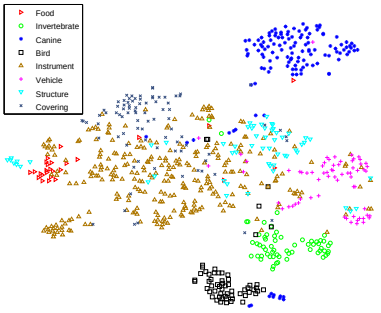


Fig. 2: An example of semantic manifold: The class prototypes of object classes from the ImageNet 2012 1K dataset are grouped into eight superclasses (food, invertebrate, canine, bird, instrument, vehicle, structure and covering) according to [13] and visualised by the 1,000D word2vec embedding [42] in a 2D space using t-SNE [38].

In this work, we explore the semantic manifold structure of class prototypes distributed in an embedding space and define a new *semantic manifold distance* for ZSL. Our approach is motivated by the inadequacies of Euclidean or cosine distance elaborated above and the fact that the distribution of class prototypes in the semantic embedding space usually has a rich semantic manifold structure. In particular, visual object classes often form groups or superclasses and the object classes from the same super-class lie on the same sub-manifold. Such a structure is illustrated clearly in Fig. 2. With the existence of such manifold structure, it is natural to conjecture that a more optimal distance would be a manifold distance which takes into account the distribution of class prototypes. The advantage of using a

1. DAP [34] and PST [49] are notable exceptions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2737007, IEEE Transactions on Pattern Analysis and Machine Intelligence

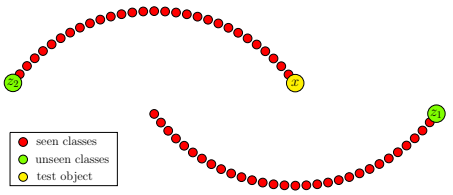


Fig. 3: The advantage of semantic manifold distance: Without considering the distribution of class prototypes captured by the semantic manifold structure, a test image x is classified as an unseen class z_1 using the Euclidean distance. When considering the prototype distribution, x is classified to z_2 by being on the same class manifold in the semantic embedding space.

semantic manifold distance over Euclidean distance is illustrated in Fig. 3. By providing a more meaningful similarity measure to the unseen class prototypes, such a semantic distance also offers a solution to the hubness and domain shift problems. For the former, the distance is computed over a manifold which reduces the dimensionality therefore hubness; for the latter, being biased towards the seen class prototypes is less detrimental as the entire manifold defined by all prototypes is used to compute the distance.

However, computing a semantic manifold distance for class prototypes is non-trivial. Specifically, different from class data samples, there is only one prototype per class; they are thus sparse, in relation to the dimensionality of the embedding space. This renders most explicit manifold learning methods unsuitable as they assume dense data distributions. In the work, we propose to model the manifold structure implicitly using a *semantic class prototype graph* where each prototype is a graph node and the connectivity on the graph is determined by the semantic relatedness between classes. ZSL is thus cast into a distance computation problem on the semantic graph. Our class prototype graph consists of two different types of graph nodes, i.e., seen and unseen prototypes, which should be treated differently. However, existing graph-based distance methods do not distinguish different nodes. To overcome this problem, we propose a new semantic distance metric on the graph based on an Absorbing Markov chain Process (AMP) specifically designed for ZSL, in which seen class prototypes are viewed as the transient states whilst unseen class prototypes the absorbing states. A test image is connected to a set of seen class nodes as a transient state on the graph, which is achieved by either a n-way seen class classifier (SR like) or the semantic embedding of the visual feature vector (SE like) as shown in Fig. 1(c). For measuring the semantic similarity distance between the image and any unseen class on the semantic graph, the Markov chain process starts from the test image node (transient node) and ends (absorbed) in one of the absorbing states (unseen class nodes). The absorbing probabilities from the test image node to unseen class prototypes are treated as the final semantic manifold distances between them.

The proposed AMP semantic distance has a number of attractive characteristics: (1) It has a closed-form solution that is very efficient to compute. (2) Importantly it is now straightforward to combine the SE and SR approaches as well as different embedding spaces. This is useful to combat the domain shift problem because as mentioned earlier each approach is susceptible to one type of domain shift but not the other. To further alleviate the domain shift problem, we introduce a regularisation term in the feature-to-semantic space project/embedding model to project an object feature vector into the semantic embedding space. The objective of the embedding model is based on a max-margin ranking loss as in [22], [1] with a new regularisation term that requires a visual sample from a seen class not only to project tightly around its seen class prototype, but also to have the chance of being close to semantically related unseen prototypes. By doing so, when the learned projection function is applied to an unseen class image, it is less likely to be biased towards the seen class prototypes. This embedding model is closely related to the AMP distance described above in that ZSL is performed by first projecting visual features into the embedding space using the proposed embedding model, followed by AMP-based recognition in that space.

Our contributions are: (1) To overcome the limitation of existing ZSL methods from relying on simplistic distance metric for NN search, we model the semantic embedding space by a rich manifold structure represented by a semantic class prototype graph. (2) A novel semantic manifold distance is formulated by exploring an Absorbing Markov chain Process (AMP) on the semantic graph, which leads to a closed-form highly efficient ZSL algorithm. (3) A new embedding model is introduced to incorporate unseen class prototypes therefore alleviate the domain shift problem in ZSL. (4) Given the new AMP model, existing SE and SR-based approaches are readily combined to complement each other. Extensive experiments on the widely used Animal with Attribute (AwA) dataset [35], the CUB-200-2011 Birds (CUB) dataset [63], the aPascal-aYahoo (aP&Y) dataset [19], and the large-scale ImageNet dataset [13] show that the proposed method outperforms significantly the state-of-the-art.

A preliminary version of this work was presented in [26]. In contrast [26], this work adds (1) an unseen prototype regularised semantic embedding (UPR-SE) model; (2) a detailed analysis of various manifold-based distances for ZSL; (3) additional evaluations on CUB and aPascal-aYahoo datasets; (4) additional discussion and evaluation on hubness problem in ZSL; (5) an additional generalised zero-shot learning experiment; and (6) new n-shot learning experiments.

2 Related Work

Semantic embedding space: Various semantic embedding spaces have been employed for zero-shot visual recognition. Earlier works used primarily semantic attributes [34], [19]. Given a defined attribute ontology, each class name is embedded in to an attribute space as a binary attribute vector.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2737007, IEEE Transactions on Pattern Analysis and Machine Intelligence

More recently, embedding based on semantic word vector space has started to gain popularity especially in large-scale zero-shot learning [22], [44], [59], [23]. Better scalability is typically the motivation as no manually defined ontology is required and any class name can be embedded for free (vs. costly labelling of attributes and ontology thereof). Beyond semantic attribute or word vector [17], [36] proposed directly learning from the rich textual descriptions of categories, such as Wikipedia articles, for ZSL. Reported results [2], [65], [35], [3], [52] seem to suggest that (1) attribute space is the most effective space which is hardly surprising as additional attribute annotations are required; and (2) combining attribute with word vector spaces often leads to improved performance. Both spaces are exploited in this work. Besides the semantic attribute and word vector spaces, context-based embedding is another popular semantic embedding model [39], in which the co-occurrence statistics of visual concepts in images is exploited for knowledge transfer. A context-based embedding is usually more robust than the embedding in a semantic space such as attribute or word vector spaces.

Projection to embedding space: Given an embedding space, most existing approaches (SE-based) also differ in the projection functions used for embedding feature vectors and can be categorised into two groups: (1) learning a projection function by regression with pre-extracted features [46], [35], [31] or end-to-end deep neural network regression [59], [22], [44]; or (2) implicitly learning the relationship between the visual and semantic spaces through a common intermediate space [36], [2], [52], [23]. The SE part of our model is based on direct projection/regression. Specifically, our project function uses a max-margin ranking loss; it is thus related to those in [22], [2], [52], [65]. What distinguishes our projection function from the existing ones is the introduction of the novel regularisation term to prevent the projected unseen class data to be biased towards the seen class prototypes in order to alleviate the projection domain shift problem.

The domain shift problem: The projection domain shift problem in ZSL was first identified by Fu et al. [24]. In order to overcome this problem, a transductive multi-view embedding framework was proposed together with label propagation on graph which requires the access of all test data at once. This assumption is often invalid in the context of ZSL because new classes typically appear dynamically and recognition needs to be done immediately. Similar transdutive approaches are proposed in [49], [31]. Instead of relying on accessing the test unseen class data as a whole by transductive learning, we tackle the domain shift problem using the proposed semantic distance as well as the new embedding model, neither of which requires the availability of the complete unseen test data set as in [24], [49], [31]. This makes our method more generally applicable in practice.

The hubness problem: The phenomenon of the presence of 'universal' neighbours, or hubs, in a high-dimensional space for nearest neighbour search was first studied by Radovanovic et al. [47]. They show that hubness is an inherent property of data distributions in a high-dimensional vector space, and a specific aspect of the curse of dimensionality. A couple of recent studies [15], [57] noted that SE-based zero-shot learning methods suffer from the hubness problem and proposed solutions to mitigate this problem. Among them, the method in [15] relies on the modelling of the global distribution of test unseen data ranks w.r.t. each class prototypes to ease the hubness problem. It is thus transductive. In contrast the method in [57] is inductive: It argued that least square regularised projection functions make the hubness problem worse and proposed to perform reverse regression, i.e., embedding class prototypes into the low-level feature space. In our work, a ranking loss is adopted to learn the projection function, to avoid the unwanted hubness-worsening property of least square-based losses. In addition, the hubness is further mitigated by computing a semantic distance instead of a simple Euclidean or cosine distance to exploit the rich manifold structure of class prototype distributions.

Manifold learning: Our ZSL model is based on a new semantic manifold distance defined on the class prototype graph in the semantic embedding space. It is thus relevant to manifold learning, a well-studied field with many models proposed including linear models (such as principal components analysis (PCA) [28] and multidimensional scaling (MDS) [56]), and nonlinear models (such as Isomap [61], locally linear embedding (LLE) [53] and Laplacian Eigenmaps [6]). Most of these models learn a manifold space explicitly where a simple Euclidean distance is computed. However, in the context of ZSL, the sparse class prototypes and high-dimensional embedding space make these conventional manifold learning models inappropriate. More relevant to our semantic distance are the distance metrics computed on a discrete graph without explicit manifold space computation. These include the shortest path distance [21] and diffusion maps distance [33]. However, not designed for ZSL, they are unable to distinguish different types of nodes (transient and absorbing nodes in our case) corresponding seen and unseen class prototypes respectively. Furthermore, our distance considers all possible paths probabilistically on the graph using a random walk process which is particularly suitable for sparse graphs at hand. More detailed analysis (Sec. 3.4) and experimental evaluations (Sec. 4.3) on the advantages of the proposed manifold distance are provided later.

Label relationship on graph: We should point out that the idea of exploiting the class label relationship as a graph is not entirely new, e.g., the WordNet has been exploited widely for transfer learning in visual recognition [51]. More recently, a specific type of label relation graph, the Hierarchy and Exclusion (HEX) graph [12] was employed for large-scale visual recognition tasks including ZSL. The HEX is a hierarchical graph of class labels, while our semantic graph is an graph of class prototypes in a semantic embedding space, designed for representing the manifold structure in that space. [18] is another relevant work, in which the image distance is measured through embedding in a semantic manifold. However, in [18], the semantic

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2737007, IEEE Transactions on Pattern Analysis and Machine Intelligence

(image) manifold is constructed using the labelled training images, while in this work, the semantic (class) manifold is constructed only using the class prototypes in a semantic embedding space.

3 ZSL Semantic Manifold Distance

3.1 Problem Definition

Let $\mathcal{Y} = \{y_1, \ldots, y_p\}$ denote 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 $X_{\mathcal{Y}} = \{(x_j, y_j)\}$ where 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 $X_{\mathcal{Z}} = \{(x_i, z_i)\}$ is provided where x_i is a d-dimensional feature vector extracted from the i-th unlabelled test image and the unknown $z_i \in \mathcal{Z}$. The goal of zero-shot learning is to learn a classifier $f : X_{\mathcal{Z}} \rightarrow \mathcal{Z}$ to predict the unseen class label z_i.

3.2 Unseen Prototype Regularised Semantic Embedding (UPR-SE)

The first step of a ZSL method is to choose a semantic embedding space. This space is used for two purposes: (1) To measure the distance between an embedded test image and an unseen class prototype in a semantic embedding (SE) based method, and (2) to measure the semantic relatedness between different classes by computing a distance between their corresponding prototypes in a semantic relatedness (SR) based method. In this work, two of the most widely used spaces are considered: attribute space and semantic word vector space. For an attribute space, a manually defined attribute ontology is required, with which each class label is represented in the attribute space (its dimension is the number of attributes) as an attribute vector. For a word vector space, similar to [59], [22], [23], [2], [64], we adopt the skip-gram text model introduced in [41], [42]. This language sentence model learns from a large text corpus to represent each English word or bi-gram (class name in the context of ZSL) as a fixed-length continuous embedding vector, so that semantically related words (e.g. horse and zebra) are adjacent in this embedding space. For notation conciseness, we denote the semantic embedding vector or class prototype of a class label y_j as $\bar{y}_j$, regardless which embedding space is used.

Next, if a SE approach is taken, an embedding model is required to project an object feature vector to a semantic vector in the semantic embedding space. The proposed semantic embedding model, termed as Unseen Prototype Regularised Semantic Embedding (UPR-SE), adds a domain shift repellent regularisation term to a max-margin ranking loss formulation. Margin-based ranking loss has been used in structured SVMs [62], [45] and recently employed for learning a ZSL visual feature embedding model [2], [22]. With a standard ranking loss, the embedding function is a linear transformation with a trainable parameter matrix M from a visual feature space to a semantic space, i.e., for a visual feature x, its embedding in the semantic space is Mx. As in [2], [22], [1], the dot-product similarity in the semantic embedding space is applied and for a class prototype $\bar{y}$, its similarity with respect to the embedding of x is $\bar{y}^T Mx$. During training, the ranking loss objective requires the correct label/prototype to be ranked higher than any of wrong prototypes. This learning objective essentially aims to push the projection of a seen training image feature vector to be close to the prototype of its corresponding seen class as well as being simultaneously far away from all other seen prototypes, as illustrated in Fig. 4(a). Concretely, for a pair of training data (x_j, y_j), the ranking loss objective is defined as:

(a) Ranking loss embedding (b) UPR-SE

Fig. 4: (a) The conventional ranking loss objective will force the projected visual feature vectors (yellow circle) from the same seen class to be tightly around their corresponding prototype in the semantic embedding space. (b) By considering the unseen class prototypes, our UPR-SE model will generalise better from seen to unseen classes.

$$loss(x_j, y_j) = \sum_{y_k \neq y_j} \max[0, l(x_j, y_j, y_k)], \qquad (1)$$

where

$$l(x_j, y_j, y_k) = margin - \bar{y}_j^T Mx_j + \bar{y}_k^T Mx_j. \qquad (2)$$

In this work the $margin$ is set to be 1 and 0.1 for the attribute and word vector space respectively.

If the objective of learning the embedding model M was to recognise test seen class data, this standard ranking loss makes sense: it will project each seen class image tightly around its corresponding seen class prototype. However, the objective of ZSL is to use this embedding model to project the unseen class data points to be close to their (unknown) unseen class prototypes. Since those unseen class prototypes were not considered in the embedding model in Eq. (1), there is no guarantee that this will happen. In fact, as shown in [24], the projected unseen class data points are often biased towards some seen class prototypes and far away from the unseen class prototypes they belong, resulting in poor recognition performance. To rectify this problem and importantly to do it in an inductive manner, we propose to use the unseen class prototypes to regularise the ranking loss objective.

More specifically, we introduce an additional regularisation term to the standard ranking loss in Eq. (1). As shown in Fig. 4(b), with this additional regularisation term, our new learning objective requires that if an unseen class z_r

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2737007, IEEE Transactions on Pattern Analysis and Machine Intelligence

is semantically related (i.e., has a small Euclidean distance in the embedding space[2]) to a seen class y_j, the embedding of a seen class data point x_j from y_j should be close to both the prototype vectors of y_j and z_r. The intuition is that if during testing, an unseen class data point of z_r is projected using the embedding model, it will not be pulled too close to y_j, i.e., being biased. Instead, it will have a better chance to be close to the correct prototype z_r rather than some arbitrary hubs. Formally, the loss function for UPR-SE is:

$$loss(x_j, y_j) = \sum_{y_k \neq y_j} \max[0, l(x_j, y_j, y_k)] + \lambda \sum_{z_r \in \mathcal{N}_{y_j}} w_{jr} [\bar{y}_j^T M x_j - \bar{z}_r^T M x_j]^2, \quad (3)$$

where $\mathcal{N}_{y_j}$ is a prototype set consisting of K neighbouring unseen class prototypes of the seen class prototype $\bar{y}_j$, and w_{jr} is the similarity/distance between $\bar{y}_j$ and z_r used to weight the pulling power of each unseen prototype in this neighbourhood for the projection Mx_j. The regularisation term is weighted by λ which is set to $\lambda = 0.1$ in this work. Our final embedding model M is learned by minimising the loss objective in Eq. (3), through Stochastic Gradient Descent (SGD).

3.3 Absorbing Markov Chain Process (AMP)

We propose to measure the distance/similarity between a projected unseen class data point and an unseen class prototype using a semantic manifold distance. To represent the manifold structure of the distribution of class prototypes, we first construct a class prototype graph. Such a graph is essentially a nearest neighbour graph, that is, on the graph, each class prototype (regardless seen or unseen) will have a corresponding graph node. This node is connected with a set of K_1 other class prototype nodes that correspond to the most semantically related classes. Again the semantic relatedness/similarity between classes is measured using the Euclidean distance between their prototypes in the semantic embedding space. Note, in this graph, the unseen class prototype nodes are only connected to the seen class prototype nodes, with reasons to be explained below. Each edge connecting two graph nodes has a weight w_{ij} which is computed out of the Euclidean distance between the two nodes in the embedding space.

To compute the distance between an unseen class data point and an unseen class prototype, we define an absorbing Markov chain process on the class prototype graph. More specifically, each seen class prototype node is viewed as a *transient* state and each unseen class prototype node an *absorbing* state, whilst the transition probability from node i to node j is computed as $p_{ij} = w_{ij} / \sum_j w_{ij}$, i.e. the normalised similarity. An absorbing state means that for each unseen class prototype node i, we set $p_{ii} = 1$ and $p_{ij} = 0$ for $i \neq j$. Note that since all of the unseen class nodes are absorbing states, any path generated by the absorbing Markov chain process will not include more than one unseen class node.

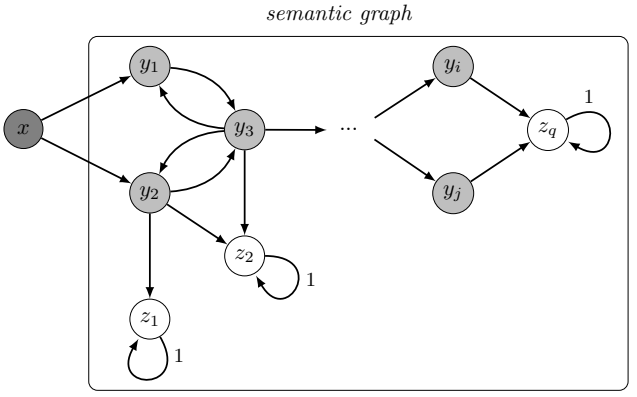


Fig. 5: After incorporating a test image into a semantic class prototype graph, zero-shot learning can be viewed as an extended absorbing Markov chain process (AMP) on the graph.

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 defined on the class prototype graph has the following canonical form:

$$P = \left(\begin{array}{c|c} Q_{p\times p} & R_{p\times q} \\ \hline \mathbf{0}_{q\times p} & \mathrm{I}_{q\times q} \end{array} \right). \quad (4)$$

In Eq. (4), $Q_{p\times p}$ 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 zero-shot learning, i.e., predicting the label z_i of an unseen test image represented as a feature vector x_i, we first need to incorporate/ingest x_i into the class prototype graph. This is followed by applying an extended absorbing Markov chain process (see Fig. 5). Specifically, x_i is connected with a subset of K_2 seen class nodes[3] selected in two ways, depending on whether a semantic relatedness (SR) strategy or a visual feature semantic embedding (SE) strategy is adopted. More concretely, if a SR strategy is taken, we utilise the training dataset $X_{\mathcal{Y}}$ to learn a n-way probabilistic classifier in the visual feature space for seen classes. For a test image $x_i \notin X_{\mathcal{Y}}$, the classifier can provide a probability $p_r(y_j|x_i)$ of image x_i belonging to the seen class y_j. If a SE strategy is adopted, the test image x_i is projected into the embedding space using the proposed UPR-SE model (Sec. 3.2), and the seen class nodes with

2. Note that we assume that Euclidean distance is sufficient for measuring semantic relatedness between two prototypes but inadequate for that between visual feature embedding and a prototype especially unseen prototype due to hubness and domain shift problems explained earlier, hence the proposed semantic distance on prototype graph.

3. This means that each Markov chain process always starts from x_i, goes through a number of seen class prototypes and end up in an unseen class prototype.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2737007, IEEE Transactions on Pattern Analysis and Machine Intelligence

the highest similarities are selected. More precisely, the similarity between the embedding of x_i, i.e. Mx_i, and the prototype of the seen class j, $\bar{y}_j$ can be computed as $s_{ij} = \bar{y}_j^T M x_i$. The similarities of the edges connecting the seen classes and the test image are then normalised as a probability $p_e(y_j|x_i) = s_{ij} / \sum_j s_{ij}$ and used to select the K_2 seen class prototypes to connect. In addition, with the proposed semantic class prototype graph, the two strategies can be easily combined by simply averaging the probability p_r from semantic relatedness and the probability p_e from semantic embedding, which gives $p_c = (p_r + p_e)/2$. Given the probabilities, we have $T_i = [t_{ij}]_{1\times p}$ as a row vector of p elements. Each element is $t_{ij} = p(y_j|x_i)$ which can be computed using either p_r, p_e or p_c depending on whether a SR, SE, or SR+SE strategy is adopted.

Note that each test image x_i is incorporated into the semantic graph as a transient state. Specifically, for x_i, there is no stepping in probabilities and the Markov process can only step out from x_i to other seen class nodes. The stepping out probabilities from x_i to seen class nodes are T_i, which are the probabilities computed using the seen class classifier scores or embedding similarities as described above. Now the transition matrix $\tilde{P}$ of the extended absorbing Markov chain process has the following canonical form:

$$\tilde{P} = \left(\begin{array}{cc|c} Q_{p\times p} & \mathbf{0}_{p\times 1} & R_{p\times q} \\ (T_i)_{1\times p} & 0_{1\times 1} & \mathbf{0}_{1\times q} \\ \hline \multicolumn{2}{c|}{\mathbf{0}_{q\times(p+1)}} & \mathrm{I}_{q\times q} \end{array} \right). \tag{5}$$

In the meantime, the extended transition matrix on all transient states, including all seen class nodes and one extra test image node x_i, are written as

$$\tilde{Q}_{(p+1)\times(p+1)} = \left(\begin{array}{cc} Q_{p\times p} & \mathbf{0}_{p\times 1} \\ (T_i)_{1\times p} & 0_{1\times 1} \end{array} \right), \tag{6}$$

and the extended transition matrix between transient states and absorbing states is

$$\tilde{R}_{(p+1)\times q} = \left(\begin{array}{c} R_{p\times q} \\ \mathbf{0}_{1\times q} \end{array} \right). \tag{7}$$

Our semantic manifold distance is computed as the absorbing probability from x_i to z_j. The intuition is that if the test image x_i belongs to an unseen class z_j, it should be connected to a number of semantically related seen class prototypes. Being semantically related, there should exist some short paths between the seen class prototypes and the unseen prototype $\bar{z}_j$ following the Markov chain process, resulting in high absorbing probability or low manifold distance. Of course none of these is certain: x_i could be connected to a wrong seen class prototype; part of the manifold structure could be badly represented in the graph due to the sparseness of the nodes. However, since we are taking a global approach, allowing multiple entry points for x_i and exhausting all the possible paths to compute a global distance using the entire manifold structure, the proposed semantic distance is robust against the imperfections of either the graph construction or the ingestion of the test images. Further discussion on this in the context of alternative manifold learning models will be presented later.

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 [30]. The absorbing probability matrix $\tilde{B} = [b_{ij}]_{(p+1)\times q}$ can be computed as follows:

$$\tilde{B} = \tilde{N} \times \tilde{R}, \tag{8}$$

in which $\tilde{N}$ is the fundamental matrix of the extended absorbing Markov chain process and is defined as follows:

$$\tilde{N}_{(p+1)\times(p+1)} = (\mathrm{I} - \tilde{Q})^{-1} = \left(\begin{array}{cc} \mathrm{I}_{p\times p} - Q_{p\times p} & \mathbf{0}_{p\times 1} \\ -(T_i)_{1\times p} & 1 \end{array} \right)^{-1}. \tag{9}$$

We use the following block matrix inversion formula [27] to compute $\tilde{N}$.

$$\left(\begin{array}{cc} A & B \\ C & D \end{array} \right)^{-1} = \left(\begin{array}{cc} E & F \\ G & H \end{array} \right), \tag{10}$$

in which we have

$$\begin{cases} G = -(D - CA^{-1}B)^{-1} CA^{-1} \\ H = (D - CA^{-1}B)^{-1}. \end{cases} \tag{11}$$

Since we only care about the absorbing probabilities for the absorbing Markov chain process starting from the test image node x_i, we only need to compute the last row of $\tilde{B}$, denoted as $\tilde{B}_{p+1,\cdot}$ for x_i (x_i corresponds to the last transient state in the extended canonical form in Eq. (5)). In particular, we can apply the above block matrix inversion formula to compute the last row of $\tilde{N}$ first as

$$\tilde{N}_{(p+1),\cdot} = \left(\begin{array}{cc} (T_i)(I - Q)^{-1}, & 1 \end{array} \right)_{1\times(p+1)} \tag{12}$$

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

$$\tilde{B}_{p+1,\cdot} = (\tilde{N}_{(p+1),\cdot}) \times \tilde{R} = T_i \times (I - Q)^{-1} R. \tag{13}$$

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 x_i. If we stack the results of all test images together, we have the final matrix S as follows:

$$S = T(I - Q)^{-1} R. \tag{14}$$

In Eq. (14), T is a $n\times p$ matrix and $(I-Q)^{-1}R$ is a $p\times q$ matrix that is only related to the semantic graph structure and can be pre-computed. The only dimension variable in Eq. (14) is the number of test images n. Therefore, our method is linear with respect to the number of test images. Moreover, since the seen class number p and unseen class number q are usually much smaller than the instance number, the matrix $(I - Q)^{-1}R$ can be computed very efficiently and computed only once.

Finally, for the test image x_i, we assign it to the unseen label that has the maximum absorbing probability when

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2737007, IEEE Transactions on Pattern Analysis and Machine Intelligence

Algorithm 1: Semantic manifold distance based on absorbing Markov chain process (AMP) for ZSL

Input: The seen/unseen prototypes and a test data x_i.
Output: The label of x_i.

1 Construct the transition matrix Q and R respectively;
2 Compute the transition probabilities T_i from x_i to the *seen* class prototypes;
3 Compute the absorbing probabilities

$$S_i = T_i(I - Q)^{-1}R$$

from x_i to the *unseen* class prototypes;
4 Choose the unseen class as the label of x_i with the highest absorbing probability as in Eq. (15).

the absorbing Markov chain starts from x_i. Our final ZSL classifier is

$$f(x_i) = \arg\max_{z_j} S_{i,j} \tag{15}$$

Note, although we use the graph-based formulation, unlike [24], [49], [31] our AMP distance model is not a transductive method. Once the class label graph is constructed, it is fixed and used in the subsequent zero-shot classification process. Consequently, we only need to access a single test image to perform recognition. Our ZSL algorithm is summarised in Algorithm 1.

3.4 Alternative Manifold-based Distances

Numerous manifold learning models have been proposed in the literature. In this section, we discuss why the proposed semantic distance computed on a class prototype graph in an embedding space using the absorbing Markov chain process (AMP) is advantageous over the alternative models for ZSL.

Existing manifold-based distances can be roughly categorised into two groups:

Explicit manifold space learning Most manifold learning models belong to this group, which construct explicitly a low-dimensional semantic manifold space where standard Euclidean distance can then be deployed. A large variety of models exist, which can be either linear (e.g. principal components analysis (PCA) [28]) or nonlinear (e.g. Isomap [61], locally linear embedding (LLE) [53] and Laplacian Eigenmaps [5], [6]), and differ in whether the local (e.g. LLE and Laplacian Eigenmaps) or global (e.g. PCA and Isomap) manifold structure is to be preserved in the manifold space. With explicit dimensionality reduction these models naturally alleviate the hubness problem. However, there is a serious problem when they are applied to the ZSL problem: Instead of using data samples to learn the manifold space, the input to these models are class prototypes in a semantic embedding space. Consequently we have a handful of data points in a high dimensional space. None of the existing explicit manifold space learning models are designed for this sparse data setting and all of them would therefore struggle as validated in our experiments (see Sec. 4.3).

Manifold distance on graph Alternatively one could model the manifold structure implicitly using a data graph and define a manifold structure on the graph. This group of methods obviously are more closely related to the proposed AMP distance. The most popular graph-based manifold distance is the shortest path distance (SPD) [61]. SPD aims to compute a manifold distance by approximating the geodesic distance using the shortest distance on the graph. Several algorithms can be applied to compute the shortest distance including the Floyd's algorithm [21] and the Dijkstra's algorithm [14]. In contrast to our AMP-based distance, the main shortcoming of the SPD distance is that it only considers one possible path between a test image and each unseen class prototype, whist our AMP distance computes all possible paths exhaustively and combines them in a probabilistic manner. Our distance is thus much more robust against noise or errors incurred by either the process of ingesting a test image into the graph (visual feature embedding) or the process of constructing the label prototype graph (class label embedding). By exploiting the manifold structure globally and probabilistically, The AMP distance is also less susceptible to the hubness problem. This shortcoming of SPD is partially addressed by existing global graph distances such as diffusion maps distance (DD) [11], [33] which also considers all possible paths. Specifically, diffusion maps distance defines a distance family through a Markov chain process on graph [11], [33] and can provide a multi-scale (long-term) analysis to the graph structure through the time (scale) parameter. However, similar to SPD, diffusion maps distance is not designed for ZSL, specifically not for the extended label prototype graph where the seen class prototypes and unseen class prototypes have different meanings in the context of ZSL and thus play different roles in computing the distance (i.e. unseen class prototypes are absorbing states and always terminate the Markov process). As a result these alternative graph-based manifold distances lead to inferior performance compared to the proposed AMP-based distance (see Sec. 4.3).

4 Experiments

4.1 Datasets and Settings

Datasets: We use four datasets for our evaluations. The **Animals with Attributes** (**AwA**) dataset[4] was introduced by Lampert et al. [34], [35]. It consists of 50 classes of animals (30,475 images), and 85 associated class-level attributes. The AwA dataset also provides a pre-defined seen/unseen split for ZSL with 6,180 images of 10 classes held out for testing and the rest as seen classes for training. The same split is used in our evaluation for fair comparisons against published results. The **CUB-200-2011 Birds** (**CUB**) [63] contains 11,788 images of 200 fine-grained bird species. We use the same split as in [2] with 150 classes for training and 50 disjoint classes for testing. The **aPascal-aYahoo** (**aP&Y**) [19][5] consists of a 12,695-image subset of the

4. http://attributes.kyb.tuebingen.mpg.de/
5. http://vision.cs.uiuc.edu/attributes/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2737007, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1: A summary of the four datasets

Dataset	AwA	CUB	aP&Y	ImageNet
# Classes	50	200	32	1,000
# Images	30,475	11,788	15,339	1.2 million
# Attributes	85	312	64	–
# word2vec dimension	100	–	–	1,000

PASCAL VOC 2008 dataset[6] and 2,644 images that were collected using the Yahoo image search engine. The class sets in the PASCAL part (20 classes) and in the Yahoo part (12 classes) are disjoint, which makes them ideal for zero-shot learning. As in most previous works, the PASCAL part is used as training data, and the Yahoo part as test data. In aP&Y, 64 binary attributes are annotated at instance-level and they are transformed to class-level attribute vectors through averaging the instance-level annotations from each class. Compared to AwA and CUB, since the seen and unseen class data are from two different datasets, aP&Y provides an additional challenge of the cross-dataset bias.

AwA, CUB and aP&Y are all widely used in existing ZSL works. However, they are not really large-scale thus somewhat contradictory to the original motivation of ZSL for scaling up visual recognition. We thus select the **ImageNet** dataset [13] as the fourth dataset. In particular, we use the ImageNet 2010 1K dataset, which consists of 1,000 categories and more than 1.2 million images. We use the same training/test (seen/unseen) split as [40], [22] for fair comparison, which gives 800 classes for training and 200 classes for testing. We summarise the characteristics of the four datasets in Table 1.

Visual features: Earlier ZSL works used hand-crafted feature representations for objects. They have been replaced by deep Convolutional Neural Network (CNN) extracted features in the past two years. CNN features are thus used in our experiments for all four datasets. In order to better compare with published results, different CNN models are used in our experiments on different datasets for feature extraction. Specifically, on AwA and aP&Y, given that all recent works are tested using either VGG-19 (4096D) [58] or GoogleNet (1024D) [60] CNN features, we report our results using the same VGG-19 and GoogleNet features on AwA and aP&Y. On CUB, the GoogleNet (1024D) feature is adopted due to its advantages in ZSL over other CNN features [2], [9]. On the ImageNet dataset, AlexNet [32] is adopted for fair comparison because all published results on this dataset used this CNN model. More specifically, we trained the AlexNet from scratch using 800 seen classes. After training, for each test image, the 4,096 dimensional top-layer hidden unit activations (fc7) of the CNN are used as the features.

Semantic embedding space: For AwA, both attribute space and word vector space are used as the semantic embedding space. For the word vector space, we train the skip-gram text model to obtain the word2vec space[7] [42], [41] on a corpus of 4.6M Wikipedia documents. As for the dimensionality of the obtained word2vec space, we set it to 100 for AwA in order to compare with the recent results in [23], [2], [31]. For CUB and aP&Y, only the attribute space is used. For ImageNet, there are no attribute definitions, so only word vector space can be used. With much more classes, 100D is not sufficient; we thus adopt an 1000D word2vec space as in [40], [22], [50].

Parameters settings: There are a number of free parameters in the proposed UPR-SE model and the AMP-based semantic distance. For learning the UPR-SE embedding, we use Stochastic Gradient Descent (SGD) with the step parameter set to 0.05 on all four datasets. The regularisation term in our UPR-SE model is computed over a neighbourhood of size K (Sec. 3.2). Similarly when we construct the semantic graph, two more neighbourhood sizes need to be determined: Each seen/unseen class prototype is connected to K_1 nearest neighbours, and a given test image is ingested into the graph by connecting to K_2 seen class prototypes (Sec. 3.3). Since the ZSL problem has no validation set available (the train/test labels are disjoint), we use 20% of the seen classes in the training sets as validation sets and perform a 5-fold cross-validation to choose the optimal values of K, K_1 and K_2. Finally, when the semantic relatedness (SR) strategy is adopted, a n-way seen class classifier needs to be learned from the training data. A linear SVM classifier is used in the experiments.

TABLE 2: Evaluation on AwA in classification accuracy (%). Different types of CNN features are used: F_D for decaf [16], F_O for overfeat [55], F_V for VGG-19 [58] and F_G for GoogleNet [60] (* indicates the transductive methods).

Method	F	SI	Result
Deng et al. [12]	F_D	A	44.2
HAP [29]	F_D	A	45.6
Kodirov et al. [31]*	F_O	A+W	75.6
TMV-BLP [24]*	F_O	A+W	80.5
SS-Voc [25]	F_O	A	78.3
DAP [35]	F_V / F_G	A	57.2 / 60.1
ESZSL [52]	F_V / F_G	A	75.3 / 76.3
SSE-ReLU [65]	F_V	A	76.3
MLZSC [8]	F_V	A	77.3
JLSE [66]*	F_V	A	80.5
DeViSE [22]	F_G	A	59.0
Socher et al. [59]	F_G	A	60.8
ConSE [44]	F_G	A	63.3
RRZSL [57]	F_G	A	66.4
Ba et al. [36]	F_G	A	69.3
SJE [2]	F_G	A+W+H	73.9
HAT [3]	F_G	A	74.9
Xian et al. [64]	F_G	A+W+H	76.1
SynC^{struct} [9]	F_G	A+W	76.3
Deep-SCoRe [43]	F_G	A+W	78.3
Ours	F_V / F_G	A+W	82.9 / **86.5**

4.2 Comparison to the State-of-the-Art

4.2.1 *Evaluation on AwA*

Competitors: For AwA, we select 20 representative ZSL methods for comparison with an emphasis on the most recent and competitive methods, as shown in Table 2. These 20 models differ in various aspects: (1) Features (F):

6. http://www.pascal-network.org/challenges/VOC/.
7. https://code.google.com/p/word2vec/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2737007, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 3: Evaluation on CUB in classification accuracy (%).

Method	F	SI	Result
DAP [35]	F_G	A	36.7
DeViSE [22]	F_G	A	33.5
ConSE [44]	F_G	A	36.2
RRZSL [57]	F_G	A	45.4
ESZSL [52]	F_G	A	47.2
SJE [2]	F_G	A	50.1
DS-SJE [48]	F_G	A	50.4
SynC^{struct} [9]	F_G	A	**54.4**
Ours	F_G	A	50.2

TABLE 4: Evaluation on aP&Y in classification accuracy (%).

Method	F	SI	Result
DAP [35]	F_G	A	35.5
ESZSL [52]	F_G	A	38.2
RRZSL [57]	F_G	A	38.8
HAT [3]	F_G	A	45.4
Long et al. [37]	F_V	A	42.3
SSE-ReLU [65]	F_V	A	46.2
JLSE [66]	F_V	A	50.4
MLZSC [8]	F_V	A	53.2
Ours	F_V / F_G	A	62.0 / **63.3**

Although the AwA dataset provides the low-level features, recently the CNN features have been used [12], [2], [65], [24], [31]. The state-of-the-art performance of ZSL reported so far was mostly achieved by either using VGG-19 [58] or GoogleNet [60] features. (2) Side information (SI): This refers to what semantic information extracted from human knowledge is used. In addition to embedding each class label into either an attribute space (A) or word vector space (W), the Wordnet hierarchy (H) is used in [1], [2], [64]. (3) Most of them are based on the SE approach, with the exception being ConSE [44] which uses the SR strategy. (4) Three compared methods including TMV-BLP [23], Kodirov et al. [31] and JLSE [66] are transductive thus require all test data to be used as a whole for inference, which gives them an advantage over the other inductive learning-based models including ours.

Comparison: From the results in Table 2, we can make the following observations: (1) Our method outperforms all 20 compared methods. Compared to the inductive learning-based methods, our model beats the closest competitor Deep-SCoRe [43] by 8.2%. (2) As expected, the three transductive methods (Kodirov et al. [31], JLSE [66] and TMV-BLP [24]) are very competitive. However our method still yields superior performance despite of using less information for being inductive. (3) Many of the compared methods (e.g. [2], [52], [65]) focus on learning advanced embedding models. However, once the test image of an unseen class is projected into the embedded space, nearest neighbour search based on Euclidean distance is applied. In contrast, our model explores the semantic manifold structure in the semantic embedding space and replaces the suboptimal Euclidean distance with the semantic manifold distance computed on a class prototype graph, leading to better performance. (4) Note that our results are obtained using a manifold modelled by 50 class prototypes in AwA, which is clearly insufficient to capture the rich intrinsic structure of a semantic embedding space. Yet the result shows that our manifold distance remains effective in this embedding space sparsely populated with class prototypes.

It should be noted that the results in Table 2 are evaluated in mean *per-image* accuracy. However, as the image samples per class on AwA are imbalanced, a mean *per-class* accuracy metric is more appropriate. We compare the performance our model against that of six alternative models with publicly available codes using the mean per-class accuracy and found that the same set of conclusions can be drawn. Detailed results on all datasets and a discussion can be found in the Supplementary Material.

4.2.2 Evaluation on CUB

Competitors: Eight existing ZSL models are compared with our model on the CUB dataset. All the methods are evaluated using the same GoogleNet feature (F_G) and the semantic attribute space (A) for fair comparison.

Comparison: The results in Table 3 show that our approach outperforms DAP [35], DeViSE [22], ConSE [44], RRZSL [57] and ESZSL [52]. It yields very similar performance as SJE [2] and DS-SJE [48] but is slightly inferior to SynC^{struct} (50.2% vs. 54.4%). Importantly, these results show clearly the advantage of using the AMP distance over the Euclidean distance-based methods, such as DeViSE. It is also evident that given the initial low performance of both the SE method DeViSE and the SR method ConSE, by combining these two strategies in the proposed semantic manifold distance model, it improves significantly the performance.

4.2.3 Evaluation on aPascal-aYahoo

Competitors: With fewer methods reporting results on aP&Y, the number of competitors available are limited. In Table 4, our method is compared against eight alternative methods: Lampert et al.'s DAP [35], Romera-Paredes and Torr's ESZSL [52], RRZSL [57], the HAT model from [3], Long et al.'s method [37], Zhang and Saligrama's semantic similarity embedding [65] (SSE-ReLU), their improved model called Joint Latent Similarity Embedding (JLSE) [66] and the recent MLZSC [8]. Note that for these experiments, since all existing methods reported results using attribute space only, our model also only uses the attribute space for fair comparison. Since four of the compared methods used GoogleNet features and three of them used VGG-19 features, we evaluate our approach using both of them on aP&Y.

Comparison: The zero-shot learning results on aP&Y are shown in Table 4. Similar observations can be made. First, after considering the semantic manifold structure, our ZSL model can achieve the state-of-the-art zero-shot learning result of 63.3% using GoogleNet, 10.1% higher than the nearest competitor MLZSC [8]. Note that with only 32 classes, the class prototype graph has even fewer nodes

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2737007, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 5: The hit@5 classification accuracy (%) of compared methods on ImageNet 2010 1K.

Method	Result
ConSE [44]	28.5
DeViSE [22]	31.8
Mensink et al. [40]	35.7
Rohrbach et al. [50]	34.8
PST [49]	34.0
ESZSL [52]	28.2
Ours	**43.3**

TABLE 6: Evaluating different manifold-based distances for ZSL (%).

Method	AwA	CUB	aP&Y	ImageNet
Euclidean	59.0	33.5	43.5	31.8
PCA	55.2	30.5	42.3	30.1
Isomap	59.2	23.3	42.3	7.9
LLE	72.4	37.3	45.4	39.1
Eigenmaps	73.7	36.5	50.7	42.5
SPD	20.0	12,6	15.2	0.9
DD	59.0	31.3	41.3	34.8
Ours	**86.5**	**50.2**	**63.3**	**43.3**

than that of AwA. However, even bigger margin is achieved using our method, further validating that the semantic distance computed using AMP is particularly effective for sparse manifold modelling. Second, compared to the results in Table 2, these results also suggest that existing ZSL methods also suffer from having fewer seen classes during training as the learned embedding model would generalise more poorly to the unseen classes.

4.2.4 Evaluation on ImageNet

Competitors: Even fewer works reported results on the large-scale ImageNet dataset. For comparison, we choose six state-of-the-art alternatives. Among them, Norouzi et al.'s convex semantic embedding ZSL (ConSE) [44] is a SR-based method. As in our method, it learns a n-way probabilistic classifier for the seen classes. The result for ConSE is based on our own implementation so the same n-way classifier is used with the same AlexNet features. In contrast, DeViSE [22] and Mensink et al.'s metric learning-based method [40] are end-to-end deep embedding models which directly project an input image into the output 1,000D word vector space with the convolutional layers of the model identical to that of AlexNet. Different from other models, PST [49] is a transductive ZSL method, which learns using the full test dataset. Finally, we also compare the Romera-Paredes and Torr's ESZSL method [52] by using the author provided code and the same features[8]. Note that we could compare with more state-of-the-art methods which provide codes. However, we found that none of them, including SSE-ReLU [65] and Kodirov et al. [31], is tractable on this large-scale dataset: On a reasonably powerful computer server with 512G memory, the codes could not run due to insufficient memory. This reveals a serious problem of many existing ZSL methods: when their embedding models have a least square-based loss, rather than a margin-based one, the computation typically involves large matrix manipulation which makes them intractable for large-scale problems.

Comparison: The performance of different methods, evaluated using the flat hit@5 classification accuracy[9] as in [40], [22], [50], is compared in Table 5. The result shows that our method clearly outperforms the state-of-the-art alternatives. This superior performance can be explained by our semantic manifold-based distance metric and the ability to combine both the semantic relatedness and semantic embedding strategies in a unified framework.

4.3 Further Analysis

Comparison to alternative manifold distances: As mentioned in Sec. 3.4, our AMP-based distance on the class prototype graph is advantageous over existing explicit manifold space learning methods and alternative graph-based manifold distances. To validate this, six representative manifold-based distances are selected together with the non-learning-based Euclidean distance for comparison. Among the six manifold distance models, four learn a manifold space explicitly followed by Euclidean distance-based NN in the learned space. They are principal components analysis (PCA) [28], Isomap [61], locally linear embedding (LLE) [53] and Laplacian Eigenmaps [5], [6]. The other two are graph-based distances including shortest path distance (SPD) [61] and diffusion maps distance (DD) [11], [33]. For fair comparison, for all compared methods, the same embedding space, visual feature representation and embedding model are used as in our method. The difference is thus only in how the manifold-based distance is computed.

From Table 6, it is clear that our class prototype graph-based manifold distance achieves significantly better performance on all four datasets. It is noted that among four manifold learning methods, the globally nonlinear method Isomap performs the worst and its performance is even worse than that of the linear manifold learning method PCA and the Euclidean distance. This is due to the fact that Isomap is based on the shortest path distance which is sensitive to the noisy connections on the semantic graph. In contrast, the two locally nonlinear manifold learning methods, i.e. LLE and Laplacian Eigenmaps, perform better than the Euclidean distance. Especially, on ImageNet with a class prototype number of 1,000, the performance of LLE and Laplacian Eigenmaps is quite competitive. However, on AwA, CUB and aP&Y with smaller number of class prototypes (50, 200 and 32 respectively), LLE and Laplacian Eigenmaps are much less effective than our AMP distance. This is expected because both LLE and Laplacian Eigenmaps need enough samples to learn a good low-dimensional semantic manifold space, while our AMP distance is computed using an absorbing Markov chain

8. Note that the kernalised version could not run on a server with 512G of memory due to the 'out of memory' issue. We thus used the linear version.

9. Each image is deemed to be classified correctly if the correct label is among the top 5 predicted labels.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2737007, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 7: ZSL results (%) obtained using AwA 40 seen classes, ImageNet 1K classes and AwA 40 plus ImageNet 1K classes to construct the semantic graph on AwA. Only the 1000D word2vec space is used for embedding due to the use of ImageNet prototypes.

	AwA 40	**ImNet 1K**	**AwA 40 + ImNet 1K**
Ours	64.0	55.9	60.9

process on the semantic graph, thus much less constrained by the sparsity of the prototype distribution.

As for the two alternative graph distances, it can be seen from Table 6 that the shortest path distance (SPD) is the worst among all compared manifold-based distances, and is even much worse than the Euclidean distance. Similar to Isomap, it is mainly because SPD only considers one possible path from a test data to an unseen class prototype and thus is vulnerable to the noisy connections on the semantic graph. In comparison, the diffusion maps distance (DD) performs better; however, it still struggles to beat the Euclidean distance and is worse than the locally nonlinear explicit manifold space methods LLE and Laplacian Eigenmaps. As analysed in Sec. 3.4, the main shortcoming of DD is that it cannot treat the seen and unseen class prototypes as different types of nodes in the graph which is important for the ZSL problem at hand: the goal is to measure the similarity between a test image and an unseen class prototype; there is thus no point continuing the random walk process once it reaches the unseen class prototype.

One may wonder if the data sparsity is the main problem for existing manifold learning methods, can we simply introduce more class prototypes into the semantic embedding space, which do not belong to either the seen and unseen classes? After all, it is free to embed arbitrary number of English words into the word2vec space used in our experiments. To find out whether it is the case, we carry out an experiment on AwA using word2vec space only and our AMP distance. We compare our original distance, computed using 40 seen class prototypes (AwA 40) with two alternatives: ImNet 1K: in this model, the 1,000 ImageNet classes are used as the seen class prototypes to ingest the test images; AwA 40 + ImageNet 1K: in this model, the 1,000 class prototypes are used to augment the original 40 seen class prototypes. Table 7 shows that introducing the additional 1K prototypes would not help. Similar results are obtained for the other six alternative manifold distances compared in Table 6. The main reason still lies with the embedding model: the projection function learned in the embedding model is trained using the 40 seen classes in AwA. Adding more seen class prototypes may enrich the manifold structure, but it will also introduce more projection domain shift problems which neutralise the benefit of having a densely populated semantic space for manifold learning.

Effectiveness of unseen prototype regularisation: Table 8 compares our model with the proposed unseen prototype regularisation term for semantic embedding (UPR-SE) (Eq. (3)) and without UPR (Eq. (1), i.e., standard ranking loss). The results show that the proposed new embedding model benefits from the regularisation term on all four datasets. This suggests that reducing the projection domain shift by regularising the project function using unseen class prototypes helps.

TABLE 8: Evaluating the unseen prototype regularisation (UPR) (%).

Method	**UPR**	**AwA**	**CUB**	**aP&Y**	**ImageNet**
Ours	without	82.1	46.5	62.9	41.0
	with	**86.5**	**50.2**	**63.3**	**43.3**

Hubness reduction: One of the motivations of the proposed semantic graph distance is to reduce hubness: with one unseen class represented by a single class prototype only, nearest neighbour (NN) search is the only option; however in a high dimensional space, any NN search would suffer from the existence of hubs: prototypes that are neighbours to many test images regardless which class they come from. We found that the hubness problem is much alleviated after our AMP-based distance is used in comparison with the conventional Euclidean distance. For example, on ImageNet, among the 200 unseen class prototypes, the worst hub appears in the top-10 neighbours of 29.1% of all test images using an Euclidean distance-based NN. After using the AMP distance, this number is reduced to 10.2%.

TABLE 9: Comparative evaluation measured in AUSUC (the higher the better) for generalised zero-shot learning on AwA.

Method	**AUSUC**
DAP [35]	0.366
IAP [35]	0.394
ConSE [44]	0.428
ESZSL [52]	0.449
SynC^{struct} [9]	0.583
Ours (NN + calibration)	0.621
Ours (SVM + threshold)	**0.683**

Generalised zero-shot learning: Another ZSL test setting emerged recently is the generalised zero-shot learning (GZSL) test setting, under which the test data set contains images from *both* seen and unseen classes. We follow the same setting of [10]. Specifically, 20% of the images from the seen classes are held out and mixed with the test images from unseen classes. As in [10], the Area Under Seen-Unseen accuracy Curve (AUSUC) is adopted as the evaluation metric. AUSUC measures how well a zero-shot learning method can trade-off between recognising images from seen classes and that of unseen classes.

Two strategies are applied to our AMP approach for GZSL: (1) NN+calibration and (2) SVM+threshold. In the NN+calibration strategy, the initial GZSL result is given by a nearest neighbour classifier. Such classification scores are then calibrated per [10]. For a test image classified as unseen classes, our AMP model is further deployed to re-classify the image into one of the unseen classes. In the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2737007, IEEE Transactions on Pattern Analysis and Machine Intelligence

SVM+threshold strategy, a n-way seen class classifier based on SVM is first used to classify the test images into the seen classes. After thresholding the seen class SVM scores, those test images with scores below the threshold will be further re-classified into unseen classes using our AMP model. In this experiment, the AMP model is compared against five alternatives on AwA and the results are shown in Table 9. It is evident that our model significantly outperforms the competitors, more so with the SVM+threshold strategy.

It should be pointed out that in both strategies, if a test data is classified to an unseen class, we assume that it must belong to one of a fixed set of unseen classes. Such an assumption is unrealistic in a practical application scenario, even though it is made by almost all existing ZSL methods. A more generalised ZSL setting would thus be considering a much larger pool of unseen classes labels, most of which have no corresponding test data samples. Developing solutions to GZSL under this more generalised setting is beyond the scope of this paper and part of our ongoing work.

More experimental results can be found in the Supplementary Material document, where a n-shot learning evaluation is given, comparative results in the mean per-class accuracy are presented, the effectiveness of combining SR and SE is evaluated, the computational cost of the proposed model is reported, and some qualitative results are also included.

5 Conclusion

We have introduced a novel zero-shot learning approach based on measuring a manifold distance between a test image and an unseen class prototype on a semantic class prototype graph. This approach is designed to overcome the hubness and domain shift problems suffered by existing ZSL methods by exploiting the manifold structure of the class prototype distribution in a semantic embedding space. The sparsity problem of the distribution is overcome by introducing a novel absorbing Markov chain process for computing a manifold distance directly on the graph rather than explicitly learning the manifold space. The proposed model also has the advantage of enabling easy fusion of existing semantic relatedness (SR) based and semantic embedding (SE) based approaches for ZSL. Extensive experiments have been carried out to demonstrate that our method outperforms the state-of-the-art methods for ZSL on four benchmarks. Ongoing work includes developing a deep end-to-end embedding model that is regularised by unseen class prototypes which can be further extended to integrate the learning of the semantic embedding space (word space) also as part of the model.

Acknowledgement

The authors were funded in part by the European Research Council under the FP7 Project SUNNY (grant agreement no. 313243).

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013. 3, 5, 10
[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015. 4, 5, 8, 9, 10
[3] Z. Al-Halah and R. Stiefelhagen. How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes. In *WCACV*, 2015. 1, 4, 9, 10
[4] E. Bart and S. Ullman. Single-example learning of novel classes using representation by similarity. In *BMVC*, 2005. 2
[5] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, 2002. 8, 11
[6] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 2003. 4, 8, 11
[7] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychological Review*, 1987. 1
[8] M. Bucher, S. Herbin, and F. Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *ECCV*, 2016. 9, 10
[9] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016. 1, 9, 10, 12
[10] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, 2016. 12
[11] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 2006. 8, 11
[12] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. In *ECCV*, 2014. 4, 9, 10
[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 3, 9
[14] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1959. 8
[15] G. Dinu and M. Baroni. Improving zero-shot learning by mitigating the hubness problem. *CoRR*, 2014. 4
[16] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014. 9
[17] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*, 2013. 4
[18] C. Fang and L. Torresani. Measuring image distances via embedding in a semantic manifold. In *ECCV*, 2012. 4
[19] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1, 3, 8
[20] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2007. 1
[21] R. W. Floyd. Algorithm 97: shortest path. *Communications of the ACM*, 1962. 4, 8
[22] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 1, 3, 4, 5, 9, 10, 11
[23] 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. 1, 4, 5, 9, 10
[24] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *IEEE TPAMI*, 2015. 2, 4, 5, 8, 9, 10
[25] Y. Fu and L. Sigal. Semi-supervised vocabulary-informed learning. In *CVPR*, 2016. 9
[26] Z. Fu, T. Xiang, E. Kodirov, and S. Gong. Zero-shot object recognition by semantic manifold distance. In *CVPR*, 2015. 1, 3
[27] G. H. Golub and C. F. Van Loan. *Matrix computations*. JHU Press, 2012. 7
[28] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 1933. 4, 8, 11
[29] S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang. Learning hypergraph-regularized attribute predictors. In *CVPR*, 2015. 1, 9
[30] S. Karlin. *A first course in stochastic processes*. Academic press, 2014. 7
[31] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, 2015. 1, 4, 8, 9, 10, 11
[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 9
[33] S. Lafon and A. B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE TPAMI*, 2006. 4, 8, 11

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2737007, IEEE Transactions on Pattern Analysis and Machine Intelligence

[34] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 2, 3, 8

[35] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 2014. 3, 4, 8, 9, 10, 12

[36] J. Lei Ba, K. Swersky, S. Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*, 2015. 1, 4, 9

[37] Y. Long, L. Liu, L. Shao, F. Shen, G. Ding, and J. Han. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. In *CVPR*, 2017. 10

[38] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 2008. 2

[39] T. Mensink, E. Gavves, and C. G. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2014. 4

[40] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *ECCV*, 2012. 9, 11

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 5, 9

[42] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 2, 5, 9

[43] P. Morgado and N. Vasconcelos. Semantically consistent regularization for zero-shot recognition. In *CVPR*, 2017. 9, 10

[44] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014. 1, 4, 9, 10, 11, 12

[45] S. Nowozin and C. H. Lampert. Structured learning and prediction in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 2011. 5

[46] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009. 1, 4

[47] M. Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *JMLR*, 2010. 2, 4

[48] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016. 10

[49] M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *NIPS*, 2013. 2, 4, 8, 11

[50] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, 2011. 9, 11

[51] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where–and why? semantic relatedness for knowledge transfer. In *CVPR*, 2010. 1, 2, 4

[52] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015. 1, 4, 9, 10, 11, 12

[53] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000. 4, 8, 11

[54] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1

[55] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013. 9

[56] R. N. Shepard. Multidimensional scaling, tree-fitting, and clustering. *Science*, 1980. 4

[57] Y. Shigeto, I. Suzuki, K. Hara, M. Shimbo, and Y. Matsumoto. Ridge regression, hubness, and zero-shot learning. In *ECML PKDD*, 2015. 4, 9, 10

[58] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 9, 10

[59] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013. 1, 4, 5, 9

[60] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 9, 10

[61] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000. 4, 8, 11

[62] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 2005. 5

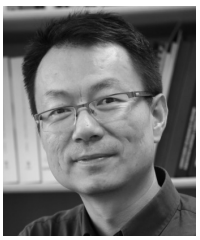
[63] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 3, 8

[64] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016. 1, 5, 9, 10

[65] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015. 4, 9, 10, 11

[66] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, 2016. 1, 9, 10

Zhenyong Fu received the Ph.D. degree in Computer Science from Shanghai Jiao Tong University in 2012. He is currently a Postdoctoral Researcher in Computer Vision Group in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests includes computer vision and machine learning.



Tao Xiang received the Ph.D. degree in electrical and computer engineering from the National University of Singapore in 2002. He is currently a reader (associate professor) in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include computer vision, machine learning, and data mining. He has published over 120 papers in international journals and conferences.

Elyor Kodirov received the Master's degree in computer science from Chonnam National University, Korea, in 2014. He is currently a Ph.D. student in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include computer vision and machine learning.

Shaogang Gong received the DPhil degree in 1989 from Keble College, Oxford University. He has been Professor of Visual Computation at Queen Mary University of London since 2001, a fellow of the Institution of Electrical Engineers and a fellow of the British Computer Society. His research interests include computer vision, machine learning, and video analysis.