Zero-Shot Object Recognition by Semantic Manifold Distance

Zhenyong Fu, Tao Xiang, Elyor Kodirov, Shaogang Gong
Queen Mary, University of London
London E1 4NS, UK
{z.fu, t.xiang, e.kodirov, s.gong}@qmul.ac.uk

Abstract

Object recognition by zero-shot learning (ZSL) aims to recognise objects without seeing any visual examples by learning knowledge transfer between seen and unseen object classes. This is typically achieved by exploring a semantic embedding space such as attribute space or semantic word vector space. In such a space, both seen and unseen class labels, as well as image features can be embedded (projected), and the similarity between them can thus be measured directly. Existing works differ in what embedding space is used and how to project the visual data into the semantic embedding space. Yet, they all measure the similarity in the space using a conventional distance metric (e.g. cosine) that does not consider the rich intrinsic structure, i.e. semantic manifold, of the semantic categories in the embedding space. In this paper we propose to model the semantic manifold in an embedding space using a semantic class label graph. The semantic manifold structure is used to redefine the distance metric in the semantic embedding space for more effective ZSL. The proposed semantic manifold distance is computed using a novel absorbing Markov chain process (AMP), which has a very efficient closed-form solution. The proposed new model improves upon and seamlessly unifies various existing ZSL algorithms. Extensive experiments on both the large scale ImageNet dataset and the widely used Animal with Attribute (AwA) dataset show that our model outperforms significantly the state-of-the-arts.

1. Introduction

Zero-shot learning (ZSL) for large scale visual object recognition has received increasing attention recently [9, 16, 23, 26, 25, 21, 17, 11]. This is because although virtually unlimited images are available via social media sharing websites such as Flickr, there are still not enough annotated images for building a model for recognising a large number of visual object classes. ZSL aims to imitate humans' ability to recognise a new class without seeing any visual

examples. A human has that ability because one is able to relate an unseen object class with the seen classes based on its semantic description. For example, assuming a child can recognise a horse; having been told that a zebra is more-or-less like a horse but with black-and-white stripes, the child has a good chance of recognising a zebra the first time it is seen. Similarly a zero-shot learning method for visual classification 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.

The seen and unseen object classes can be related in a semantic embedding space where each class label/name is represented as a high dimensional vector. The spaces used by most early works are based on semantic attributes [16]. Given a defined attribute ontology, each class name can be converted to a binary attribute vector. More recently, embedding based on semantic word space has started to gain popularity [10, 21, 29]. Better scalability is typically the motivation for this approach as no manually defined ontology is required and the space is learned using a vast unannotated text corpus by natural language processing. Such an approach can embed any class name for free (vs. costly labelling of attributes and ontology thereof). Regardless the space used, the embedded class name (a vector) is called a prototype of that class [11].

Given a semantic embedding space and a set of seen and unseen class prototypes, the semantic relatedness between an unseen class and each seen class can be measured as a distance between the two class prototypes. However, an image of a visual object is represented by a visual feature vector; its distance to the unseen class prototypes in the semantic embedding space cannot be measured directly. Existing methods for solving this problem fall into two categories. The first category (Fig. 1(a)) relies on learning a n-way discrete classifier for the seen classes in the visual feature space, which is then used to compute the visual similarity between an image of unseen class to those of the seen classes. These seen classes serve as the mediators for the unseen classes and the test images [2]. Specifically, the semantic relatedness between the seen and unseen classes is
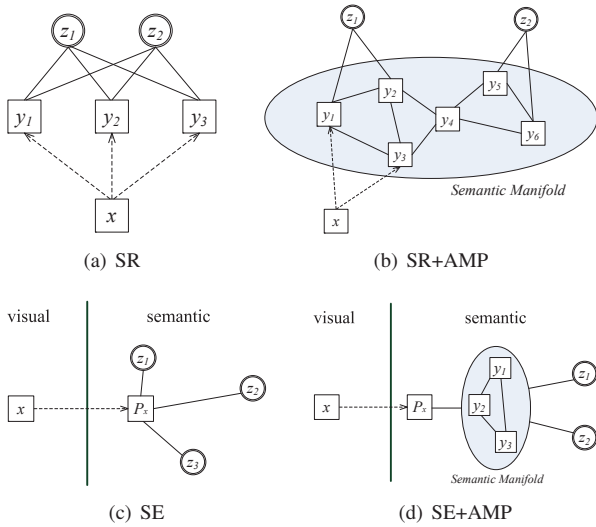
(a) SR   (b) SR+AMP

(c) SE   (d) SE+AMP

Figure 1. Our Absorbing Markov chain process (AMP) based zero-shot learning framework unifies Semantic Relatedness (SR) and Semantic Embedding (SE) 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.
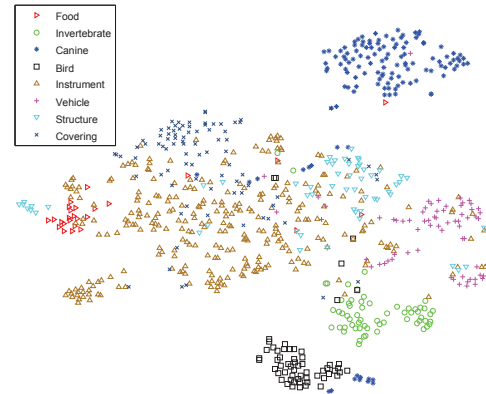

Figure 2. An example of semantic manifold. We group the classes from ImageNet 2012 1K dataset into eight superclasses (food, invertebrate, canine, bird, instrument, vehicle, structure and covering) according to [5] and visualise the 1,000D word2vec embedding [20] into 2D low-dimensional space using t-SNE [30] in 2D. It is evident that a semantic manifold structure exists and the object classes from the same superclass lie in the same manifold. In this work, we formulate a semantic manifold constrained similarity distance to solve the zero-shot learning problem.

modelled by the distance between their prototypes, or the knowledge from linguistic processing [26]. Such semantic relatedness (similarity) is then compared with the visual similarity and the image is classified to an unseen class if the two types of similarities against the mediators, *i.e.* seen classes, match. In contrast, methods in the second category (Fig. 1(c)) are based on embedding directly the visual feature vectors into a semantic embedding space [1, 10, 29]. This is typically achieved by learning a projection function between the visual feature space and the semantic embedding space. Such a function is learned from the labelled training visual data consisting of seen classes only. After this visual feature embedding (mapping) process, zero-shot classification is performed directly by measuring similarity using nearest neighbour (NN) or its probabilistic variants such as direct attribute prediction (DAP) [16]. These two categories are denoted in this paper as *Semantic Relatedness* (SR) and *Semantic Embedding* (SE) respectively.

A common characteristic of existing ZSL models from both approaches is that they all rely critically on computing the similarity distance in the semantic embedding space. All existing methods adopt a conventional distance metric computed directly in the embedding space. However, as shown in Fig. 2, the distribution of the semantic class prototypes in the semantic embedding space has a rich intrinsic manifold structure. Existing direct distance metrics ignore such structure therefore are suboptimal. In this work, we explore this semantic manifold structure in order to define a new simi-

larity distance metric between a test image and the unseen class prototypes for ZSL. We formulate a representation of this structure using a semantic graph where each class is a node and the connectivity on the graph is determined by the semantic relatedness between classes.

By exploiting the semantic manifold, we can measure a semantic distance based on two assumptions: (1) If the projection of a test image and an unseen class prototype are connected by strongly related seen class prototypes, they should be "close" (small distance or high similarity) on the manifold, and thus likely to have the same class label. We call this the *connectivity assumption*. For example, the semantic concepts grass, tree and snake are strongly related because they usually appear in the same context, *i.e.* forest, although the superclass of snake is different with that of grass and tree. If a test image and an unseen class prototype are all close to the structure (context) formed by grass, tree and snake, they are likely to have the same label, *e.g.* an animal living in forest. (2) If the projection of a test image and an unseen class prototype are on the same local structure (typically referred to as a cluster or a local manifold), they are likely to have the same label. This assumption is often called the *cluster assumption* [27, 3, 32]. For example, if a test image and an unseen class prototype fall into the same local manifold (*e.g.* bird in Fig. 2), they are likely to have the same label, *e.g.* a specific type of bird.

Based on the proposed semantic manifold representation and two assumptions above, a novel zero-shot learning (ZSL) algorithm is formulated. Specifically, given an

embedding space and a semantic graph representing the structure of the underlying semantic manifold (Figs. 1(b) and 1(d)), we first 'connect' the visual feature vector of a test image to a set of seen class nodes on the graph. This is achieved by either a n-way seen class classifier (Fig. 1(b)) or the semantic embedding of the visual feature vectors (Fig. 1(d)). For measuring the similarity distance between the image and any unseen class on the semantic manifold, we design a special Absorbing Markov chain Process (AMP), by which the seen class nodes are the transient states and the unseen class nodes are the absorbing states. Our Markov chain process starts from the test image node and ends (absorbed) in one of the absorbing states (unseen class nodes), which indicates to which unseen class this test image belongs. The proposed AMP model has a closed-form solution that is very efficient to compute. Furthermore, as shown in Figs. 1(b) and 1(d), our semantic manifold based AMP ZSL algorithm can be used in conjunction with any existing semantic relatedness or semantic embedding based ZSL method given any semantic embedding space, because different methods and spaces can be used to compute the graph connectivity and transition probabilities between nodes on the same semantic graph.

Our contributions are three-folds: (1) We propose a manifold representation of a semantic embedding space using a semantic graph of object class prototypes for exploring a richer semantic distance in ZSL. (2) A novel Absorbing Markov chain Process (AMP) is formulated on the semantic graph which leads to a closed-form efficient ZSL algorithm. (3) The proposed semantic manifold and AMP algorithm improve upon and seamlessly unify various existing ZSL learning algorithms and different semantic embedding spaces. Extensive experiments on both the large scale ImageNet dataset [5] and the widely used Animal with Attribute (AwA) dataset [15] show that our model significantly outperforms the state-of-the-arts.

2. Related Work

Existing ZSL methods differ in the semantic spaces used and how the knowledge is transferred from the seen to unseen object classes. Despite its earlier dominance, attribute based embedding spaces [16, 23, 8, 7, 1, 12] are giving away to semantic word vector based spaces [22, 29, 10, 11] due to the latter's advantage for scalability. This is because whilst the primary objective of ZSL is to solve the large scale learning problem without exhaustive labelling of data, manually defining an attribute ontology for each and every object class does not scale well.

Given a semantic space, either a visual feature semantic embedding approach or a n-way seen class classifier based semantic relatedness mapping strategy can be adopted, with the former being more popular than the latter. Examples of the semantic embedding (SE) strategy are direct attribute prediction (DAP) [16] and its variant [10]. Recently Fu *et al.* [11] pointed out that this strategy however suffers from a projection domain shift problem – the visual feature mapping (embedding) learned from the seen class data may not generalise well to the unseen class data, which is an implicit assumption for semantic embedding based ZSL. They proposed a transductive multi-view embedding framework to solve this problem. Our semantic manifold can inherently solve the projection domain shift problem. This is because we measure a manifold constrained semantic graph distance rather than a direct cosine distance between the embedded visual feature vectors and the unseen class prototypes. Critically, our method is not transductive, that is, we do not assume that the full test dataset containing unlabelled visual examples of all unseen classes is available for learning.

In contrast, the semantic relatedness (SR) based ZSL strategy has been less popular [15, 21], partly due to the task of learning a good n-way probabilistic classifier being formidable. However, recent works have reported impressive classification accuracy over 1,000 classes [14] using deep convolutional neural network learned classifiers. This advance on deep learning is removing the barrier to adopting the semantic relatedness approach to ZSL, given that such a strategy is potentially more advantageous over the semantic embedding approach [15, 21]. In this work, we provide a unified framework to enable both strategies to be combined in our AMP algorithm, resulting in an overall better model as demonstrated in our extensive experiments.

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 [26]. More recently, a specific type of label relation graph, the Hierarchy and Exclusion (HEX) graph [4] is employed for large scale visual recognition learning tasks including ZSL. The HEX is a hierarchical graph of class labels, while our semantic graph is an undirected graph of class prototypes in a semantic embedding space, designed for representing the manifold structure in that space. To our best knowledge, this work is the first attempt to explore the manifold structure of and derive a semantic graph distance for a semantic embedding space. Our experiments show that our model significantly outperforms the HEX graph model of [4] on the ZSL task.

3. Methodology
3.1. Problem Definition

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

and the unknown $y_i \in \mathcal{Z}$. The goal of ZSL is to learn a classifier $f : X \to \mathcal{Z}$ to predict y_i.

3.2. Semantic Embedding Space

For any ZSL method, the similarity or semantic relatedness between seen and unseen classes needs to be computed. This is typically achieved by a semantic embedding space. 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). An attribute vector is denoted as $\mathbf{y}_j^A$. For a word vector space, similar to [29, 10, 11], we adopt the skip-gram text model introduced in [19, 20]. This language sentence model learns from a large text corpus to represent each English word (and bi-gram) as a fixed-length continuous embedding vector $\mathbf{y}_j^V$, so that semantically related words (*e.g.* horse and zebra) are adjacent in this embedding space.

The semantic space is used for two purposes in a ZSL learning framework: (1) To measure the semantic relatedness between different classes by computing a distance between their corresponding prototypes, and (2) to measure the semantic similarity between a test image and a class prototype. For this purpose, the visual feature vector $\mathbf{x}_i$ needs to be projected into the semantic space and represented as $\mathbf{x}_i^A$ or $\mathbf{x}_i^V$ depending on which embedding space is used. This projection can be realised by classification [15] or regression [29, 10, 11].

3.3. Semantic Graph

Next we describe how to represent the manifold structure of a semantic embedding space by constructing a graph. A semantic graph is constructed as a k-nearest-neighbour graph using the seen and unseen class prototypes. On the semantic graph, each class prototype (regardless seen or unseen) will have a corresponding graph node which is connected with its k most similar (semantically related) other classes. This definition of similarity is based on a distance between two class prototypes in the semantic embedding space. Note, the unseen class nodes are only connected with the seen class nodes. The reason is explained below (Sec. 3.4). The edge weight w_{ij} of the semantic graph is the similarity between two end nodes of an edge. More details about the semantic graph construction are given in Sec. 4.

3.4. Absorbing Markov Chain Process

We define an absorbing Markov chain process on the semantic graph as follows. Each unseen class node is viewed as an *absorbing* state and each seen class node is viewed as a *transient* state, whilst the transition probability from class node i to class node j is $p_{ij} = w_{ij} / \sum_j w_{ij}$, *i.e.* the normalised similarity. An absorbing state means that for each unseen class node i, we have $p_{ii} = 1$ and $p_{ij} = 0$ for $i \neq j$. Note that since all of the unseen class nodes are absorbing
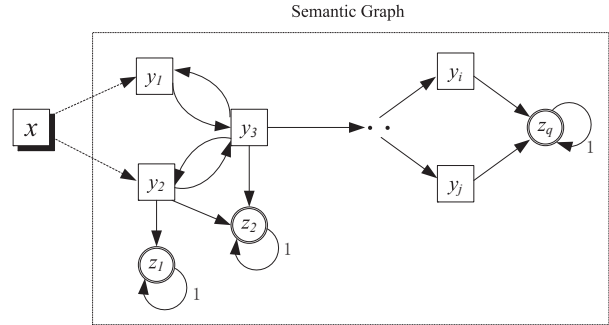

Figure 3. After incorporating a test image into a semantic graph, zero-shot learning can be viewed as an extended absorbing Markov chain process (AMP) on the graph.

states, any path generated by the absorbing Markov chain process will not include more than one unseen class node.

We re-number the class nodes (as states in a Markov process) so that the seen class nodes (transient states) come first. Then, the transition matrix P of the above absorbing Markov chain process has the following canonical form:

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

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

3.5. Zero-shot Classification

For zero-shot learning, *i.e.* predicting the label y_i of an unseen test image $\mathbf{x}_i$, we first need to incorporate $\mathbf{x}_i$ into the semantic graph. This is followed by applying an extended absorbing Markov chain process (see Fig. 3). In order to incorporate a test image $\mathbf{x}_i$ into the semantic graph, it is connected with a selected set of K seen class nodes. There are two ways by which the seen class nodes are selected for connection, depending on whether a n-way seen class classifier plus semantic relatedness (SR) strategy or a visual feature semantic embedding (SE) strategy is adopted (Sec. 2). More specifically, if the former 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 $\mathbf{x}_i \notin X_{\mathcal{Y}}$, the classifier can provide a probability $p_r(y_j | \mathbf{x}_i)$ of image $\mathbf{x}_i$ belonging to the seen class y_j. If the second strategy is adopted, the test image $\mathbf{x}_i$ is projected into the embedding space and becomes $\mathbf{x}_i^A$ or $\mathbf{x}_i^V$ (Sec. 3.2) and the seen class nodes with the K smallest distance are selected. More precisely, the similarity between $\mathbf{x}_i^A$ or $\mathbf{x}_i^V$ and the

prototype of the seen class j, $\mathbf{y}_j^A$ or $\mathbf{y}_j^V$ can be computed as s_{ij}. Then we normalise the similarity as the probability $p_e(y_j|\mathbf{x}_i) = s_{ij}/\sum_j s_{ij}$. The node representing image $\mathbf{x}_i$ is then connected to the seen classes with the K highest probabilities. In addition, our framework combines these two strategies by 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 $t_{ij} = p(y_j|\mathbf{x}_i)$ 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.

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

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

In the meanwhile, the extended transition matrix on all transient states, including all seen class nodes and one extra test image node $\mathbf{x}_i$, are written as

$$\tilde{Q}_{(p+1) \times (p+1)} = \left(\begin{array}{cc} Q_{p \times p} & \mathbf{0}_{p \times 1} \\ (T_i)_{1 \times p} & 0_{1 \times 1} \end{array} \right), \qquad (3)$$

and the extended transition matrix between transient states and absorbing states should be

$$\tilde{R}_{(p+1) \times q} = \left(\begin{array}{c} R_{p \times q} \\ \mathbf{0}_{1 \times q} \end{array} \right). \qquad (4)$$

In the extended semantic graph, it is obvious that if the test image $\mathbf{x}_i$ is close to one unseen class node, $e.g.$ z_j, on the semantic graph, the absorbing Markov chain process that starts from $\mathbf{x}_i$ will have a high probability to be absorbed at z_j. Thus, the probability of $\mathbf{x}_i$ being labelled as the unseen class label represented by z_j should be high. Note that this absorbing probability is determined solely by the structure of the semantic manifold.

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

$$\tilde{B} = \tilde{N} \times \tilde{R}, \qquad (5)$$

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

$$\tilde{N}_{(p+1) \times (p+1)} = (I-\tilde{Q})^{-1} = \left(\begin{array}{cc} I_{p \times p} - Q_{p \times p} & \mathbf{0}_{p \times 1} \\ -(T_i)_{1 \times p} & 1 \end{array} \right)^{-1}. \qquad (6)$$

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

$$\left(\begin{array}{cc} A & B \\ C & D \end{array} \right)^{-1} = \left(\begin{array}{cc} E & F \\ G & H \end{array} \right), \qquad (7)$$

in which we have

$$\begin{cases} G = -(D - CA^{-1}B)^{-1}CA^{-1} \\ H = (D - CA^{-1}B)^{-1}. \end{cases} \qquad (8)$$

Since we only care about the absorbing probabilities for the absorbing chain process starting from the test image node $\mathbf{x}_i$, we only need to compute the last row of $\tilde{B}$, denoted as $\tilde{B}_{p+1,\cdot}$ for $\mathbf{x}_i$ ($\mathbf{x}_i$ corresponds to the last transient state in the extended canonical form in Eq. (2)). In particular, we can apply the above block matrix inversion formula to compute the last row of $\tilde{N}$ as

$$\tilde{N}_{(p+1),\cdot} = \left(\begin{array}{cc} (T_i)(I - Q)^{-1}, & 1 \end{array} \right)_{1 \times (p+1)} \qquad (9)$$

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

$$\tilde{B}_{p+1,\cdot} = (\tilde{N}_{(p+1),\cdot}) \times \tilde{R} = T_i \times (I - Q)^{-1}R. \qquad (10)$$

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

$$S = T(I - Q)^{-1}R. \qquad (11)$$

In Eq. (11), 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. (11) 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 $\mathbf{x}_i$, we assign it to the unseen label that has the maximum absorbing probability when the absorbing chain starts from $\mathbf{x}_i$. Finally, our ZSL classifer is

$$f(\mathbf{x}_i) = \arg\max_{z_j} S_{i,j} \qquad (12)$$

Note, although we use the graph based formulation, unlike [11] our AMP method is not a transductive method.

The semantic graph in our approach is only related to the seen/unseen class prototypes. Once the semantic graph is constructed, it is fixed and used in the subsequent zero-shot learning process. In addition, it is noted in [11] that multiple semantic embedding spaces contain complementary information thus should be combined for ZSL. This can be easily achieved using AMP by averaging the similarity matrices obtained on different spaces.

4. Experiments

4.1. Datasets and Settings

Datasets. Two datasets are chosen for our evaluations, ImageNet and AwA. ImageNet [5] is a large scale image dataset suitable for ZSL evaluation. 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 [18, 10] for fair comparison, which gives 800 classes for training and 200 classes for testing. Only a handful of previous works report results on ImageNet, thus limiting our comparison. Therefore the AwA (animals with attributes) dataset [15] is selected as the second dataset, on which the majority of ZSL models proposed so far have been tested. AwA provides 50 classes of animals (30,475 images), and 85 associated class-level attributes. Different from ImageNet, both attribute space and semantic word space can be evaluated using the AwA dataset. AwA also provides a defined seen/unseen split for ZSL with 10 classes and 6180 images held out as in [16].

Visual Features. On the ImageNet dataset, we pre-train a deep convolutional neural network (CNN) using the training dataset with 800 classes, following the model architecture in [14]. After training, for each test image, the 4,096 dimensional top-layer hidden unit activations (fc7) of the CNN are taken as the features. On AwA, we also use the CNN feature originally provided [28] as it has been shown recently to be much more powerful than the low-level features originally provided in [15].

Semantic Embedding Space. For the semantic embedding space, semantic word vector space is used for both datasets. We train the skip-gram text model [20, 19] on a corpus of 4.6M Wikipedia documents to form a 1000-D and a 100-D word spaces for the ImageNet 2010 and AwA datasets respectively. In addition, for AwA, each class label is represented as an 85D attribute vector in the attribute space. The mapping/embedding of visual feature vector (4,096D) into the 1000/100D word vector space is achieved using the deep CNN model DeViSE [10]. On ImageNet 2010, we set the $margin = 0.1$ as in [10], and on AwA, we set the $margin = 1$. For learning the deep DeViSE model, we use Stochastic Gradient Descent (SGD) with the step parameter set to 0.05 as in [10] on both ImageNet and AwA. When the semantic relatedness strategy is adopted, a n-way seen class classifier needs to be learned from the training data.

We use the Liblinear toolbox [6] to train a L_2-regularised multi-class logistic regression classifier as in [16].

Semantic Graph. We use the k-nearest-neighbour to set up the semantic graph (Sec. 3.4). At first, the seen class prototypes are used to set up a semantic subgraph, in which we use $k = 10$ on ImageNet and $k = 2$ and $k = 3$ respectively for attribute and word2vec semantic space on AwA. Then, the unseen classes are connected into the seen semantic subgraph and each unseen class is connected to its k-nearest seen class prototypes, in which we set $k = 20$ on ImageNet and $k = 8$ and $k = 4$ respectively for attribute and word vector semantic space on AwA. For the attribute and word vector prototypes, we compute the cosine similarity as the edge weights. Finally, each test image is connected to K nearest seen classes (Sec. 3.5). We set $K = 10$ for ImageNet and $K = 4$ and $K = 10$ respectively for attribute and word vector semantic space on AwA. The effects of varying the values of these free parameters will be evaluated in Sec. 4.4.

4.2. Evaluations on ImageNet

Method	Result
ConSE [21]	28.5%
DeViSE [10]	31.8%
Mensink *et al.* [18]	35.7%
Rohrbach *et al.* [25]	34.8%
PST [24]	34.0%
Our AMP (SR+SE)	**41.0%**

Table 1. The hit@5 classification accuracy of compared methods on ImageNet 2010 1K.

Competitors. Our method is compared against five state-of-the-arts alternatives. They are either semantic relatedness (SR) based or semantic embedding (SE) based, while our method is based on a combination of semantic relatedness and embedding (SR+SE). More specifically, Norouzi *et al.*'s convex semantic embedding ZSL (ConSE) [21] is SR based. As in our method, it learns a n-way probabilistic classifier for the seen classes. The results for ConSE is based on our own implementation so the same n-way classifier is used. In contrast, DeViSE [10] and Mensink *et al.*'s metric learning based method [18] project the 4,096 CNN features into the 1,000D word vector space. Like [11], PST is a transductive ZSL method, which learns using the full test dataset. In contrast, other four methods, including our AMP, only use only the training dataset for model learning. **Comparative Results.** The performance of different methods, evaluated using the flat hit@5 classification accuracy[1] as in [18, 10, 25], is compared in Table 1. The result shows that our method clearly outperforms the state-of-the-art al-

---

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

	Our AMP	Nearest Neighbor		Our AMP	Nearest Neighbor
	banana pineapple, ananas mashed potato quince coffee bean	broom **banana** pineapple, ananas hare zucchini, courgette		**cranberry** pine, pine tree grape walnut linden	persimmon Japanese pagoda fragrant orchid chrysanthemum American white birch
	minivan motor scooter, scooter horse cart, horse-cart fireboat alp	cannon **minivan** bullet train, bullet web site, website home theater		**motor scooter, scooter** horse cart, horse-cart go-kart minivan carousel, carrousel	sling, scarf bandage gasmask, respirator **motor scooter, scooter** chain saw, chainsaw crutch
	dalmatian, coach dog **Brittany spaniel** Eskimo dog, husky Staffordshire bullterrier griffon, Brussels griffon	dalmatian, coach dog dingo, warrigal fur coat **Brittany spaniel** dogsled, dog sled		**sea anemone, anemone** nematode starfish, sea star sea urchin Japanese pagoda tree	sea slug, nudibranch Japanese pagoda tree celandine poppy cosmos, cosmea **sea anemone, anemone**
	desktop computer jigsaw puzzle joystick baseball **photocopier**	dishwasher shaver, electric shaver ballpoint pen hand glass desktop computer		**hand glass** shower cap ladle magazine, mag birdhouse	dishwasher ballpoint pen **hand glass** cliff, drop, drop-off drum

Figure 4. Qualitative results on ImageNet. For each image, the top 5 zero-shot predictions of our AMP and nearest neighbour (NN) classifier, both trained on ImageNet 2010 800. Predictions are ordered by decreasing score, with correct predictions in bold.
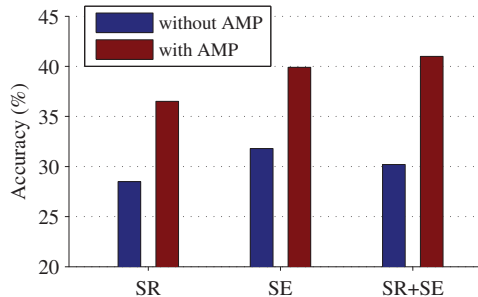

Figure 5. Evaluation of the contributions of individual novel component of our model on ImageNet.

ternatives. Some qualitative results can be seen in Fig. 4. This superior performance can be explained by our semantic manifold based distance metric algorithm and the ability to combine both the SE and SR strategies in a unified framework. Next we investigate further how each of the these two novel components contributes to the overall performance.

Contributions of Individual Novel Components. First, we compare in Fig. 5 the performance of our method with and without the AMP algorithm under semantic relatedness (SR), semantic embedding (SE) and the combination of both. Note, our SR model without AMP is equivalent

to the ConSE model, and our SE model without AMP is equivalent to the DeViSE model. It can be observed that (1) Both a semantic relatedness (*i.e.* ConSE) or semantic embedding (*i.e.* DeViSE) based method can benefit from our AMP framework. Interestingly, after incorporating our AMP, the result of SR+AMP (SR with AMP) can achieve 36.5%, which is already higher than the state-of-the-art results on ImageNet 2010, *i.e.* DeViSE's 31.8%, Mensink *et al.*'s 35.7% and Rohrbach *et al.*'s 34.8%. When we use our AMP to replace the nearest neighbour in DeViSE (based on cosine distance), the performance has an almost 10% improvement. Some qualitative results are shown in Fig. 4. It shows that compared to the cosine distance in nearest neighbour ZSL, our semantic graph distance is much more meaningful (*e.g.* not only the correct labels are predicted, closely related labels are also ranked high). (2) After we combine the SR and SE settings together, we can achieve our final result 41.0%. However, without AMP (SE+SR without AMP), the result of 30.2%, obtained by score level fusion is even worse then SE (31.8%) without AMP alone. This result suggests that the graph level fusion of both strategies is superior to the simple score level fusion which may have a negative effect. In conclusion both the semantic manifold based distance metric and the combination of SE and SR strategies contribute to our superior performance.

Method	S. Space	Feature	Result
IAP [15]	A	L/C	42.2/44.5
DAP [15]	A	L/C	41.4/53.2
DS [26]	W/A	L/C	35.7/52.7
AHLE [1]	A	L	43.5
Yu *et al*. [31]	A	L	48.3
Jayaraman *et al*. [13]	A	L	43.0
TMV-BLP [11]	A+W	L	47.1
Deng *et al*. [4]	A	L/C	38.5/44.2
Our AMP (SR+SE)	A+W	C	**66.0**

Table 2. Results on AwA in classification accuracy (%). We compare with the state-of-the-arts under different semantic embedding spaces including word vector (W) and visual attribute (A). Two types of features are used: low-level (L) and CNN (C) features.

4.3. Evaluations on AwA

Competitors. Compared to ImageNet, far more published results on AwA are available, as compared in Table 2. Apart from taking either a SR or SE based strategy, they also differ in the semantic embedding space used, as both the attribute and word vector spaces are available for AwA. Both [11] and our AMP model can exploit both spaces. However only our method is able to combine both the SR and SE strategies. These models also differ in the feature space used. The dataset provided low-level features (L) were used in most studies. However, more recently the CNN features (C) have been used [4]. Moreover, TMV-BLP [11] is transductive thus requires all test data for learning, and Yu et al [31] uses additional human annotations.

Comparative Results. Table 2 shows that the best result is obtained using the proposed AMP (SR+SE) method, with two observations: (1) In general using the CNN features leads to better performance. Given CNN features, our model outperforms significantly the other existing methods. This is partly because we use the deep CNN model directly to learn the projection rather than just extract the features. (2) Our performance is much better than that of Deng *et al*. [4] which exploits a semantic label graph. This shows that exploiting a graph based manifold modelling of the semantic embedding space is clearly more beneficial than in the label space. 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. We thus expect the result to be further improved when more seen classes are added.

Contributions of Individual Novel Components. Similar to the evaluation in the previous section on ImageNet, In Fig. 6, we evaluate the contribution of the AMP algorithm, and the combination of both SR and SE strategies. Similar conclusions can be drawn, that is, both components help and naive score level fusion of both strategies is inferior to our coherent graph based fusion.
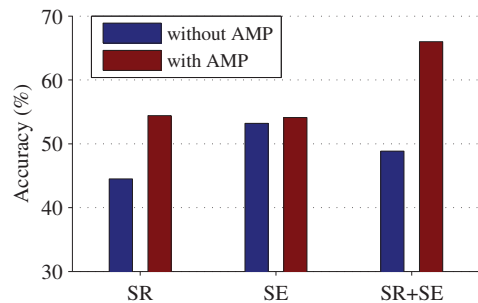

Figure 6. Evaluation of the contributions of individual novel component of our model on AwA.
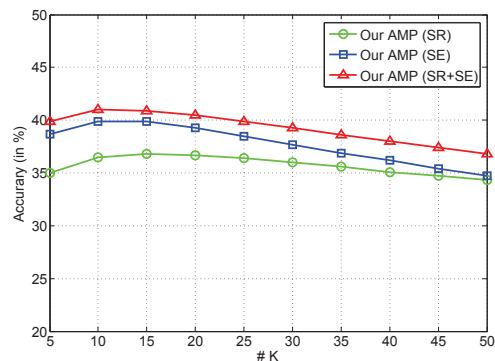
4.4. Further Evaluations

Figure 7. The performance (top-5 results in %) of our AMP methods with respect to different settings of the parameter K.

Parameter sensitivity. We evaluate the effect of setting different values of K, *i.e.* the number of top similar seen classes that a test image will connect, on ImageNet. From Fig. 7, it is evident that different versions of our method are all stable for different K value. Similar observation is made for the other free parameter k, *i.e.* how many class prototypes are connected with each node in the graph.

Running time. On a Tesla K20m GPU server, it takes on average 5.13 milliseconds to classify a single test image on ImageNet. This includes 5.08 milliseconds for mapping the image into the word space using DeViSE. Our model is thus extremely efficient.

5. Conclusion

We have introduced a novel zero-shot learning approach based on formulating a semantic manifold distance. We proposed an absorbing Markov chain process for ZSL classification with efficient closed-form solution. Importantly the proposed a framework enables seamless fusion of existing semantic relatedness based and semantic embedding based methods for ZSL. We have shown experimentally that our method outperforms the state-of-the-arts methods for ZSL on widely used benchmarks.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 819–826. IEEE, 2013. 2, 3, 8

[2] E. Bart and S. Ullman. Single-example learning of novel classes using representation by similarity. In *BMVC*, volume 1, page 2, 2005. 1

[3] O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. In *Advances in neural information processing systems*, pages 585–592, 2002. 2

[4] 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 *Computer Vision–ECCV 2014*, pages 48–64. Springer, 2014. 3, 8

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 2, 3, 6

[6] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008. 6

[7] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2352–2359. IEEE, 2010. 3

[8] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009. 3

[9] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2007. 1

[10] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013. 1, 2, 3, 4, 6

[11] 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, 3, 4, 5, 6, 8

[12] D. Jayaraman and K. Grauman. Zero shot recognition with unreliable attributes. *arXiv preprint arXiv:1409.4327*, 2014. 3

[13] D. Jayaraman and K. GraumanE. Zero shot recognition with unreliable attributes. In *NIPS*, volume 1, 2014. 8

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, volume 1, page 4, 2012. 3, 6

[15] C. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot learning of object categories. 2013. 3, 4, 6, 8

[16] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009. 1, 2, 3, 6

[17] T. Mensink, E. Gavves, and C. G. Snoek. Costa: Co-occurrence statistics for zero-shot classification. 2014. 1

[18] 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 *Computer Vision–ECCV 2012*, pages 488–501. Springer, 2012. 6

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 4, 6

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013. 2, 4, 6

[21] 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, 3, 6

[22] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014. 3

[23] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, volume 3, pages 5–2, 2009. 1, 3

[24] M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *Advances in Neural Information Processing Systems*, pages 46–54, 2013. 6

[25] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1641–1648. IEEE, 2011. 1, 6

[26] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where–and why? semantic relatedness for knowledge transfer. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 910–917. IEEE, 2010. 1, 2, 3, 8

[27] M. Seeger. Learning with labeled and unlabeled data. Technical report, 2000. 2

[28] 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. 6

[29] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, pages 935–943, 2013. 1, 2, 3, 4

[30] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008. 2

[31] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. *CVPR*, 2013. 8

[32] X. Zhu, Z. Ghahramani, J. Lafferty, et al. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 3, pages 912–919, 2003. 2