The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Towards Affordable Semantic Searching: Zero-Shot Retrieval via Dominant Attributes

Yang Long,[1,3] **Li Liu,**[†,2] **Yuming Shen,**[4] **Ling Shao**[1,2,4,*]
[1] School of Automation, Northwestern Polytechnical University, Xi'an, P. R. China
[2] JD Artificial Intelligence Research (JDAIR), Beijing, P. R. China
[3] Open Lab, School of Computing, Newcastle University, Newcastle upon Tyne, UK
[4] School of Computing Sciences, University of East Anglia, Norwich, UK

Abstract

Instance-level retrieval has become an essential paradigm to index and retrieves images from large-scale databases. Conventional instance search requires at least an example of the query image to retrieve images that contain the same object instance. Existing semantic retrieval can only search *semantically-related* images, such as those sharing the same category or a set of tags, not the exact instances. Meanwhile, the unrealistic assumption is that all categories or tags are known beforehand. Training models for these semantic concepts highly rely on instance-level attributes or human captions which are expensive to acquire. Given the above challenges, this paper studies the *Zero-shot Retrieval* problem that aims for instance-level image search using only a few dominant attributes. The contributions are: 1) we utilise automatic word embedding to infer class-level attributes to circumvent expensive human labelling; 2) the inferred class-attributes can be extended into discriminative instance attributes through our proposed Latent Instance Attributes Discovery (LIAD) algorithm; 3) our method is not restricted to complete attribute signatures, query of dominant attributes can also be dealt with. On two benchmarks, CUB and SUN, extensive experiments demonstrate that our method can achieve promising performance for the problem. Moreover, our approach can also benefit conventional ZSL tasks.

Introduction

Conventional recognition approaches usually focus on predicting the class label of an image. Recently, instance-level retrieval has aroused increasing attention due to its unique value for real-world applications, such as person identification (Haque, Alahi, and Fei-Fei 2016), place localisation (Wu and Rehg 2008), and specific object search (Tao, Smeulders, and Chang 2015). Instead of predicting the general class label, *e.g. person* or *restaurant*, we hope to find the exact object of the query image, *i.e.* '*Who is the person?*' or '*Which restaurant is shown?*' from a large number of candidates. Previous work has achieved excellent results on instance retrieval within a single category, such as buildings (Arandjelovic and Zisserman 2012), logos (Tao et al. 2014), and book covers (Wang et al. 2011). However, developing generalised instance retrieval across different classes remains an open problem.

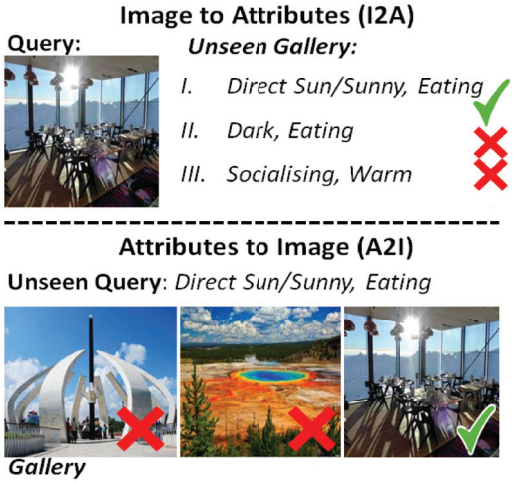

Figure 1: In the zero-shot retrieval problem, there are two situations: 1) querying image from the unseen gallery where only several spare attributes of each instance are available; 2) retrieving an unseen query in the gallery by given dominant attributes.

There are mainly three practical issues that are worth considering. Firstly, it has been estimated that humans can distinguish at least 30,000 object classes (Biederman 1987). For generalised instance retrieval, it is infeasible to collect labelled data to train an retrieval model for each class separately (Tao, Smeulders, and Chang 2015). Secondly, an unknown query instance is not guaranteed to have a class name. In other words, many class-specific models may not be applicable for generalised retrieval tasks. Thirdly, many recent approaches are based on one-shot learning techniques, such as person re-identification (Liao et al. 2015), and face verification (Taigman et al. 2014). However, in many realistic applications, *e.g.* criminal identification, even acquiring one query image could be unattainable.

Inspired by the fact that humans can identify an unseen instance given only some dominant attributes, this paper defines a more flexible scenario named *Zero-shot Retrieval*

* The Corresponding Author is Ling Shao.
† This author contributes equally to the work.

(ZSR): given training images from disjointed seen classes, we aim to search an unseen instance given a query of a set of dominant attributes. For example, a generalised retrieval machine may not know whether the query image comes from people or other classes. One may wish to find an unseen criminal from the gallery of mixed classes where only verbal descriptions from the witnesses are available as the query. There are two possible scenarios illustrated in Fig. 1. In the **Image to Attributes I2A (I2A)** scenario, the task is to find the query image from unknown classes in the gallery where only dominant attributes of each instance are available. Inversely, in the **Attributes to Image (A2I)** scenario, given a query of dominant attributes, the task is to answer which image in the gallery is queried.

Related work. Instance Retrieval is different from canonical recognition tasks or fine-grained classifications since the expected outputs are the identical instances to the query in the gallery rather than its class label. In conventional retrieval approaches (Arandjelovic and Zisserman 2012; Tao et al. 2014; Wu and Rehg 2008), the training set, gallery, and query images are restricted to the same class. As shown in (Tao, Smeulders, and Chang 2015), a model for a particular class cannot be directly applied to other classes. To break such a restriction, cross-class instance retrieval can be achieved by first precisely predicting the class label and then zooming into the instance-level (Tao, Smeulders, and Chang 2015). However, in many realistic scenarios, sufficient training samples cannot be guaranteed for each class. Thus, one shot learning based approaches (Fei-Fei, Fergus, and Perona 2006) are proposed and are successfully employed to address the re-identification (Liao et al. 2015) and face verification (Taigman et al. 2014) problems. In order to overcome a wide range of inner-instance variations, semantic side information, such as attributes, is widely adopted in recent robust systems (Kumar et al. 2009; Long, Zhu, and Shao 2016; Long et al. 2017c).

Zero-shot learning techniques also share the advantages of using attributes but go one step further. Despite some performance sacrifice, ZSL can predict the class label of an unseen instance without providing even one visual instance. The core assumption of ZSL is that the learnt models can transcend the class boundaries (Long et al. 2017b; 2017a; Long and Shao 2017a), such as some basic visual terms like colours, textures, shapes, and parts. More advanced attributes are designed by ontology engineers. For example, the adopted datasets in this paper, SUN (Patterson et al. 2014) and CUB (Wah et al. 2011), provide many relevant attributes to depict scenes and birds. Such well-designed attribute models have shown promising generalisation to unseen classes (Yu and Aloimonos 2010; Parikh and Grauman 2011; Jayaraman and Grauman 2014; Jayaraman, Sha, and Grauman 2014; Lampert, Nickisch, and Harmeling 2014; Changpinyo et al. 2016; Akata et al. 2013; Long, Liu, and Shao 2017; Long and Shao 2017b; Long, Liu, and Shao 2016; Qin et al. 2017).

In order to circumvent expensive instance-level attributes, several approaches have been proposed to alleviate the cost of annotating attributes, such as text-based (Rohrbach et al. 2010; Elhoseiny, Saleh, and Elgammal 2013; Mensink, Gavves, and Snoek 2014; Qiao et al. 2016) or word-embedding methods (Socher et al. 2013; Frome et al. 2013; Norouzi et al. 2014; Al-Halah, Tapaswi, and Stiefelhagen 2016) that can automatically associate semantic models to visual appearances in a data-driven manner. However, the class-level description is still too coarse for instance retrieval.

Contributions

- Compared to conventional ZSL, expensive attribute annotations are substituted by word embedding our proposed inference algorithm. Either a class name or several dominant attributes can be simply encoded into an *Augmented Attribute* (AA) Representation.
- A Latent Instance Attributes Discovery (LIAD) algorithm is proposed to automatically extend class-level attributes into instance-level. We investigate how to use an orthogonal projection to find discriminative latent instance-attributes.
- While testing is performed, unlike previous ZSL methods using the complete attribute representation for classification, we use only a few dominant attributes to infer more instance-sensitive representations.

Zero-shot Retrieval via Dominant Attributes

During the training phase, we only use visual features and class labels of training images. In order to circumvent expensive human attributes, class labels are encoded by word embedding and converted to our proposed representation named Augmented Attributes (AA) through class-attribute inference. Hereby, each class gains a semantic representation. For the sake of flexibility, our AA is a unified representation, which can encode three types of *Concepts*: class names, attributes, and a set of dominant attributes. The former two are used in training phase and the last one is used during the test. The second step is to infer latent instance attributes by our orthogonal embedding algorithm. During the test, a query of either dominant attributes or a visual instance can be encoded by AA and then projected into the embedding space to make the retrieval. We formalise each step as follows:

I. Class-level AA Inference The training set contains extracted visual features with class labels in N pairs: $(x_1, y_1), \ldots, (x_N, y_N) \subseteq \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = [x_{nd}] \in \mathbb{R}^{N \times D}$ is a D-dimensional feature space; and $\mathcal{Y} = \{1, \ldots, C\}$ consists of C discrete class labels. For ZSR, we need to find an instance from unseen classes by their dominant attributes. Therefore, we need to define a universal set S of all involved class name and attribute concepts. We extract Word2Vec features (Mikolov et al. 2013) for each involved concept in an L-dimensional word embedding space $\mathcal{W} = [w_s] \in \mathbb{R}^{S \times L}$, where $S = C + M$ is composed of C class labels $\mathcal{W}_c = [w_c] \in \mathbb{R}^{C \times L}$ and M attribute concepts $\mathcal{W}_a = [w_m] \in \mathbb{R}^{M \times L}$. The class-level AA is achieved by $\mathcal{A} = [a_{cs}] = \Psi(w_c, w_s) \in \mathbb{R}^{C \times S}$, where Ψ is an inference function introduced later. Note that the class labels are also used as attributes for better generalisation. Such representations are named as *Augmented Attributes*. Note that this is not the same method in (Sharmanska, Quadrianto, and

Lampert 2012) that augments attributes with driven data by auto-encoder models. Our approach just adds class names into attributes to form a S-dimensional AA space for better generalisation.

II. Latent Instance-Attributes Discovery Using the inferred class-level AA as clues, we aim to discover the unique AA of each visual instance. Initially, we assume instances in a class have the same AA representations, which results in $\mathcal{A} = [a_{ns}] \in \mathbb{R}^{N \times S}$. Our key idea is to find a latent space $\mathcal{V} \in \mathbb{R}^{N \times K}$ that can remove the correlation between instances and between attributes so that the learnt representations are unique to each other. As a result, we can achieve K discriminative latent attributes that are sensitive to instances at the training phase.

III. Zero-shot Retrieval via dominant attributes During the test, both query and the gallery images come from unseen classes in $\mathcal{Z} = \{z : z \notin \mathcal{Y}\}$. There are two possible scenarios: **1) Attributes to Image (I2A)** Given a set of dominant attributes as a query, the task is to retrieve the corresponding visual instance in the image gallery; **2) Image to Attributes (I2A)** Given a query image, the task is to predict its identity in the gallery where only dominant attributes of each instance are available. For both scenarios, we first infer the full augmented attribute vector $\hat{a}$ from given dominant attributes using Ψ. Eventually, both inferred attribute vectors and visual features can be projected into the orthogonal space $\mathcal{V}$ to make the retrieval.

Augmented Attributes Inference

In this paper, a concept can be a class name, an attribute, or a set of attributes, anyone of which can consist of several words. We consider each single word is a local representation of the whole concept, *i.e.* $w = \{w^i\}$. Specifically, we extract Word2Vec features (Mikolov et al. 2013) for each word of involved class and attribute concepts. Using the extracted Word2Vec features, class-attribute relationship $(w_c, w_s) = (\{w_c^i\}, \{w_s^j\})$ is inferred as most ZSL methods, where i and j indicate the i-th and j-th words in a concept. Here, the attribute list is augmented from M to $S = C + M$ by including class concepts as well. Hereby, the relations between classes are also taken into account so as to achieve better generalisation. The relations between a pair of concepts can be inferred by:

$$\begin{aligned}\Psi(w_c, w_s) &= \Phi(w_c)^\top \Phi(w_s) \\ &= \frac{1}{|w_c||w_s|} \sum_{i=1}^{|w_c|} \sum_{j=1}^{|w_s|} \psi(w_c^i, w_s^j),\end{aligned} \tag{1}$$

where Φ and ψ are operations on concept and word levels, and $|\cdot|$ denotes the cardinality of a feature set. Φ is an embedding function that maps a set of word vectors to a high-dimensional embedding space. Eq. 1 can be computed by repeating the word-level operation ϕ:

$$\psi(w_c^i, w_s^j) = \phi(w_c^i)^\top \phi(w_s^j), \forall\, w_c^i \in w_c, \forall\, w_s^j \in w_s. \tag{2}$$

Due to a large variety of words, such as nouns, and adjectives, a direct comparison between two words may not lead to the best result. Instead, we employ the log-odds ratio $\log(\mathbb{O}(\cdot))$ (McCann and Lowe 2012) as our embedding function ϕ to estimate whether the word is likely to be correlated to other words:

$$\phi(w_c^i) = [\log\left(\mathbb{O}_{w_1}(w_c^i)\right), \ldots, \log\left(\mathbb{O}_{w_S}(w_c^i)\right)]^\top, \tag{3}$$

where $\forall\ w_s \in \{w_1, ..., w_S\}$. By assuming the prior is uniform, we apply Bayes' Rule to obtain:

$$\begin{aligned}\log\left(\mathbb{O}_{w_s}(w_c^i)\right) &= \log \frac{\mathrm{p}(w_s|w_c^i)}{\mathrm{p}(\bar{w}_s|w_c^i)} = \log \frac{\mathrm{p}(w_c^i|w_s)\,\mathrm{p}(w_s)}{\mathrm{p}(w_c^i|\bar{w}_s)\,\mathrm{p}(\bar{w}_s)} \\ &= \log \frac{\mathrm{p}(w_c^i|w_s)}{\mathrm{p}(w_c^i|\bar{w}_s)} + \log \frac{\mathrm{p}(w_s)}{\mathrm{p}(\bar{w}_s)} \\ &= \log\left(\mathrm{p}(w_c^i|w_s) - \mathrm{p}(w_c^i|\bar{w}_s)\right),\end{aligned} \tag{4}$$

where $\mathrm{p}(w_c^i|\bar{w}_s)$ denotes the likelihood that the word w_c^i is absent in w_s. Eq. 4 provides an intuitive conclusion. By using the odds ratio, only prominent words that relate to several classes will gain high likelihoods, whereas words that are too common or too rare will be suppressed. The log-likelihoods in Eq. 4 can be estimated using

$$\begin{aligned}\log\left(\mathbb{O}_{w_s}(w_c^i)\right) &= \log(\delta_i^s - \bar{\delta}_i^s), \text{ where} \\ \delta_i^s &= \exp\left(\|w_c^i - NN^s(w_c^i)\|_2\right) \\ \bar{\delta}_i^s &= \exp\left(\frac{1}{S-1} \sum_{\bar{s} \neq s} \|w_c^i - NN^{\bar{s}}(w_c^i)\|_2\right),\end{aligned} \tag{5}$$

where δ_i^s is the distance from w_c^i to its nearest neighbour belonging to concept w_s and $\bar{\delta}_i^s$ represents the average distance from w_c^i to the nearest neighbours in other concepts; $NN^s(\cdot)$ is a nearest neighbour search of all words in concept s; and $\|\cdot\|_2$ is the ℓ_2-norm distance. By repeating Eq. 1 to 5 between each pair of class label c and concepts s, we can infer class-attributes for each class:

$$a_c = [\Psi_1(w_c, w_1), \ldots, \Psi_1(w_c, w_S)]. \tag{6}$$

Latent Instance Attributes Discovery

By assuming all instances in class c with the same inferred class-attributes a_c, the training set becomes $(x_1, a_1, y_1), \ldots, (x_N, a_N, y_N) \subseteq \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$. A conventional ZSL classifier can be trained by:

$$\min_P \mathcal{L}(\mathcal{X}P - \mathcal{A}) + \lambda\Omega(P), \tag{7}$$

where $\mathcal{L}(.)$ is the loss function and Ω is the regularisation term with hyper-parameter λ. However, such a framework can only differentiate classes. Also, some dimensions of AA are highly correlated. In order to retrieve each instance rather than the class label, the key challenge is to discover the unique attributes of each visual instance that are different from the general class-attribute signature. We propose a *Latent Instance Attributes Discovery (LIAD)* algorithm to find a set of mutually independent bases to break the bias which assumes all of the class-attributes are correlated, *i.e.* for all of the instances in the same class, all attributes are always either present or absent together. The bases can be discovered by an orthogonal constrained projection:

$$\min_{P_1, P_2} \|\mathcal{X}P_1 - \mathcal{V}\|_F^2 + \gamma\|\mathcal{A}P_2 - \mathcal{V}\|_F^2, \text{ s.t. } \mathcal{V}^\top\mathcal{V} = I, \tag{8}$$

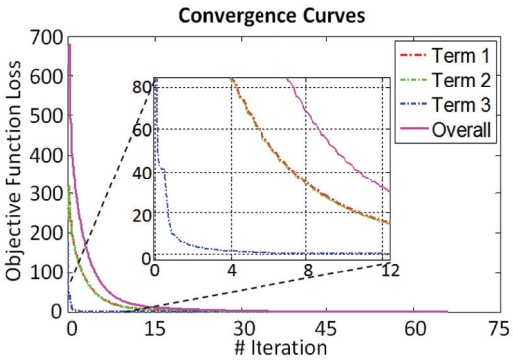

Figure 2: Convergence curves with respect to the number of iterations. Terms one and two correspond to the reconstruction error of visual and semantic spaces. Term three controls the orthogonality.

where $\|.\|_F$ is the Frobenius norm of a matrix; I is the identity matrix; $\mathcal{V} \in \mathbb{R}^{N\times K}$ is the shared space between $\mathcal{X}$ and $\mathcal{A}$; γ is a hyper-parameter that controls the balance between $\mathcal{X}$ and $\mathcal{A}$; $P_1 \in \mathbb{R}^{D\times K}$ and $P_2 \in \mathbb{R}^{S\times K}$ are two projection matrices. K is a hyper-parameter for spectral clustering, through which we aim to cluster correlated concepts of class labels or attributes into K mutually independent latent attributes. Within the same class, the discovered latent attributes are close to each other. Meanwhile, different instances can also be discriminated by the latent attributes since they are achieved by minimising the reconstruction error to each unique visual instance in $\mathcal{X}$.

Optimisation Eq. 8 is a non-convex NP-hard problem due to the orthogonal constraint. To best exploit that, we propose an alternating optimisation scheme based on coordinate descent. We first initialise $\mathcal{V}$ as a random orthogonal matrix that is s.t. $\mathcal{V}^\top\mathcal{V} = I$. The following optimisation sequentially updates P_1, P_2 and $\mathcal{V}$ using:

$\mathbf{P_1}$-step. By fixing P_2 and $\mathcal{V}$, we can reduce Eq. 8 to

$$\min_{P_1} \|\mathcal{X}P_1 - \mathcal{V}\|_F^2. \tag{9}$$

The resulting equation is derived by a standard least squares problem with the following analytical solution:

$$P_1 = (\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top\mathcal{V}. \tag{10}$$

$\mathbf{P_2}$-step. Similar to P_1-step, by fixing P_1 and $\mathcal{V}$, we can obtain P_2 as

$$P_2 = (\mathcal{A}^\top\mathcal{A})^{-1}\mathcal{A}^\top\mathcal{V}. \tag{11}$$

$\mathcal{V}$-step: By fixing P_1 and P_2, we can reshape Eq. 8 to the following problem:

$$\min_{\mathcal{V}} \|\mathcal{X}P_1 - \mathcal{V}\|_F^2 + \gamma\|\mathcal{A}P_2 - \mathcal{V}\|_F^2, \text{ s.t. } \mathcal{V}^\top\mathcal{V} = I. \tag{12}$$

To solve Eq. 12 with the orthogonality constraint, we adopt the gradient flow in (Wen and Yin 2013) which is an iterative scheme that can optimise orthogonal problems with a feasible solution. Specifically, given the orthogonal matrix $\mathcal{V}_t$ during the t-th iterative optimisation, a better solution of $\mathcal{V}_{t+1}$ is updated via *Cayley transformation*:

$$\mathcal{V}_{t+1} = H_t\mathcal{V}_t, \tag{13}$$

where H_t is the *Cayley transformation* matrix and defined as

$$H_t = \left(I + \frac{\tau}{2}\Phi_t\right)^{-1}\left(I - \frac{\tau}{2}\Phi_t\right), \tag{14}$$

where $\Phi_t = \Delta_t\mathcal{V}_t^\top - \mathcal{V}_t\Delta_t^\top$ is the skew-symmetric matrix, τ is an approximate minimiser satisfying Armijo-Wolfe conditions (Wright and Nocedal 1999) and Δ_t is the partial derivative of Eq. 12 with respect to $\mathcal{V}$ as

$$\Delta_t = 2(1+\gamma)\mathcal{V} - 2\mathcal{X}P_1 - 2\gamma\mathcal{A}P_2. \tag{15}$$

In this way, we repeat the above formulation to update $\mathcal{V}$ until achieving convergence in the inner-loop. The convergence proof of the above updating procedure with the orthogonality constraint can be found in (Wen and Yin 2013).

The overall optimisation is achieved by alternating P_1, P_2 and $\mathcal{V}$ steps up to T epochs. Generally, $T = 15 \sim 20$ is proved to be enough for convergence as shown in Fig. 2. The overall LIAD is summarised in Algorithm. 1.

Algorithm 1: LIAD

Input: Training set $\{\mathcal{X},\ \mathcal{Y}\}$, inferred $\mathcal{A}$, K, and γ
Output: P_1 and P_2

Initialise random orthogonal matrix $\mathcal{V}$, s.t. $\mathcal{V}^\top\mathcal{V} = I$.
Repeat
  P_1**-Step:** Fix $\mathcal{V}$, P_2 and update P_1 using Eq. 10.
  P_2**-Step:** Fix $\mathcal{V}$, P_1 and update P_2 using Eq. 11.
  $\mathcal{V}$**-Step:** Fix P_1, P_2 and update $\mathcal{V}$ by following steps:
  for $t = 1 : \max$ iterations **do**
    Compute the gradient Δ_t using Eq. 15;
    Compute the the skew-symmetric matrix Φ_t;
    Compute the Cayley matrix H_t using Eq. 14;
    Compute the $\mathcal{V}_{t+1}$ using Eq. 13;
    if convergence, **break**;
  end
Until convergence
Return P_1, P_2 and $\mathcal{V}$

Zero-shot Retrieval via Sparse Attributes

An unseen instance can be depicted by a set of dominant attributes. In this paper, we select a number of attributes with top confidence scores in the provided instance-attributes as dominant attributes. Using the proposed method in Section , we can infer the AA vectors of unseen instances and project them into the latent attribute space to make the retrieval. More specifically, we denote the dominant attributes in word embedding forms as $\hat{w} = \{\hat{w}^i\}$. All of the involved words in the dominant attributes are pooled together. The idea is consistent to Section that regards each single word as a local feature of the unseen instance description. We then estimate the likelihood between each augmented attribute w_s and the unseen instance $\hat{w}$ using Eq. 1 to 5, through which we can get the full attribute vector $\hat{a} = [\Psi(\hat{w}, w_1), \ldots, \Psi(\hat{w}, w_S)]$.

Using P_1 and P_2, either visual features or attribute vectors can be projected into the latent attribute space, which enables us to achieve both of the ZSR scenarios mentioned

Table 1: Main ZSR results at different ranks. Upper and lower results correspond to that of I2A and A2I scenarios. In order to show the overall trend of the outcomes, we average the hit rates in all of the classes. Best results are in bold.

	SUN Attribute					Caltech-UCSD Birds				
Method	@Rank1	@Rank5	@Rank10	@Rank20	@Rank50	@Rank1	@Rank5	@Rank10	@Rank20	@Rank50
DAP (Lampert, Nickisch, and Harmeling 2009)	7.5	18.8	34.9	48.5	61.2	3.80	5.82	12.61	17.92	24.25
ALE (Akata et al. 2013)	14.8	29.6	47.5	64.2	78.4	7.81	18.23	22.52	30.74	38.72
ESZSL (Romera-Paredes and Torr 2015)	19.9	38.8	56.2	69.7	82.8	15.28	20.34	25.88	38.21	40.72
LatEm (Xian et al. 2016)	25.3	38.4	62.8	70.1	85.2	17.42	24.82	32.48	40.96	46.81
CCA	9.2	20.5	43.3	63.0	74.8	10.16	15.30	24.79	32.56	36.16
NoV	7.2	17.5	49.2	62.3	76.3	9.83	16.42	19.50	29.82	33.78
Ours	**28.7**	**42.2**	**68.5**	**72.8**	**86.2**	**19.82**	**27.53**	**36.20**	**44.12**	**48.83**
DAP (Lampert, Nickisch, and Harmeling 2009)	8.8	19.2	32.6	44.7	52.5	5.42	8.82	14.27	16.82	22.36
ALE (Akata et al. 2013)	12.2	26.7	43.0	61.5	72.2	12.87	16.43	24.50	29.98	34.71
ESZSL (Romera-Paredes and Torr 2015)	**18.8**	34.2	49.1	66.2	76.9	14.31	17.40	23.65	36.48	39.22
LatEm (Xian et al. 2016)	17.3	36.4	58.8	67.6	**80.8**	15.82	20.26	29.48	36.25	43.82
CCA	13.8	27.4	44.5	62.8	70.7	10.43	14.56	18.85	25.58	30.76
NoV	14.1	23.0	36.8	43.4	49.6	7.47	14.42	18.68	23.72	28.80
Ours	18.7	**37.7**	**61.9**	**70.2**	78.8	**18.61**	**26.62**	**32.81**	**39.42**	**44.28**

Table 2: Key statistics of the SUN and CUB datasets

Dataset	SUN	CUB
# instances	14,340	11,788
# attributes	102	312
seen/unseen splits	707/10	150/50
attribute type	ins.+ cont.	ins.+ bin.
# total concepts	819	512
Unseen Gallery size	200	2933

earlier. **Scenario 1:** Suppose there are $\hat{N}$ instances from unseen classes $\mathcal{Z}$ in the gallery without images. Only their dominant attributes are available, which can be denoted as: $\hat{\mathcal{W}} = [\hat{w}_{\hat{n}}] \in \mathbb{R}^{\hat{N}\times L}$; and can be converted into full attribute vectors $\hat{\mathcal{A}} = [\hat{a}_{\hat{n}}] \in \mathbb{R}^{\hat{N}\times S}$ using Eq. 1 as illustrated above. During the test, a query image $\hat{x}$ can be retrieved by:

$$\hat{id} = \arg\min_{\hat{n}} \|\hat{x}P_1 - \hat{a}_{\hat{n}}P_2\|_2^2. \tag{16}$$

Scenario 2: The gallery is composed of $\hat{N}$ images from unseen classes $\mathcal{Z}$: $\hat{\mathcal{X}} = [\hat{x}_{\hat{n}}] \in \mathbb{R}^{\hat{N}\times D}$. We first project their visual features into the latent attribute space. During the test, the query is a set of dominant attributes $\hat{w} = \{\hat{w}^i\}$ depicting the requested instance, from which the full attribute vector $\hat{a}$ can be inferred using Eq. 1. The retrieval is then made by:

$$\hat{id} = \arg\min_{\hat{n}} \|\hat{x}_{\hat{n}}P_1 - \hat{a}P_2\|_2^2. \tag{17}$$

Experiments

Our method is evaluated on two benchmarks for zero-shot recognition, SUN (Patterson et al. 2014) and CUB (Wah et al. 2011). Up-to-date statistics are summarised in Table 2. All of the visual features used in our experiments are based on the *VGG-19* model (Simonyan and Zisserman 2014) that was released by (Zhang and Saligrama 2015). For word embedding, we adopt the *GoogleNews-vectors-negative300* model (Mikolov et al. 2013) that was trained on the partial Google News dataset using about 100 billion words. We follow the seen/unseen splits as that of conventional ZSL settings. However, the evaluations focus on the ZSR performance. For I2A and A2I scenarios, we alternatingly use the attributes of unseen instances and images as gallery and queries.

Evaluation Protocol We adopt the hit rate as the main evaluation protocol. Given a query, we check whether the corresponding instance is found in a number of top ranks. Since the instances are from multiple classes, to show the overall trend of the outcomes, we average the hit rates in different classes.

Implementations Our semantic inference procedures have no hyper-parameters. We cross-validate all hyper-parameters of our LIAD on the training set. Since we do not use any attributes during the training phase, we propose a 5-fold approximated cross-validation scheme for the ZSR problem. Under each set of hyper-parameters, we first achieve $\mathcal{V}$ on the whole training set. The $\mathcal{V}$ of each instance is used as the inferred outputs of dominant attributes $\mathcal{V}_{\text{attr}}$. We then randomly divide the training classes into five folds. For each fold, we compute a new pair of $P_1,\ P_2$ using the other four folds and the achieved P_1 is used to project the visual instances in the validation fold into $\mathcal{V}_{\text{vis}}$. Using the $\mathcal{V}_{\text{attr}}$ achieved on the whole training set and the $\mathcal{V}_{\text{vis}}$ projected from validation set, the retrieval performance can be validated.

Main Results

Table 1 summarises our comparison with both the state-of-the-art methods and some baselines for self-comparison. DAP (Lampert, Nickisch, and Harmeling 2009) is one of the earliest ZSL approaches that makes predictions based on Maximum Likelihood criteria. ALE (Akata et al. 2013) and ESZSL (Romera-Paredes and Torr 2015) optimise the objective functions that enforce correct labels to rank higher than incorrect ones. LatEm (Xian et al. 2016) shares the similar objective but is based on bilinear compatibility functions. For all of above methods, we implement with their released codes and follow the details as reported in the original papers. Besides, we use CCA to compute the shared space V as the baseline method. To investigate the contribution of our orthogonal regularisation, we also compute the shared space based on a linear solution without the orthogonal constraint, which is denoted as NoV.

For all of the above methods, inferred class-level AA in Section is used to train the models. During the test, we se-

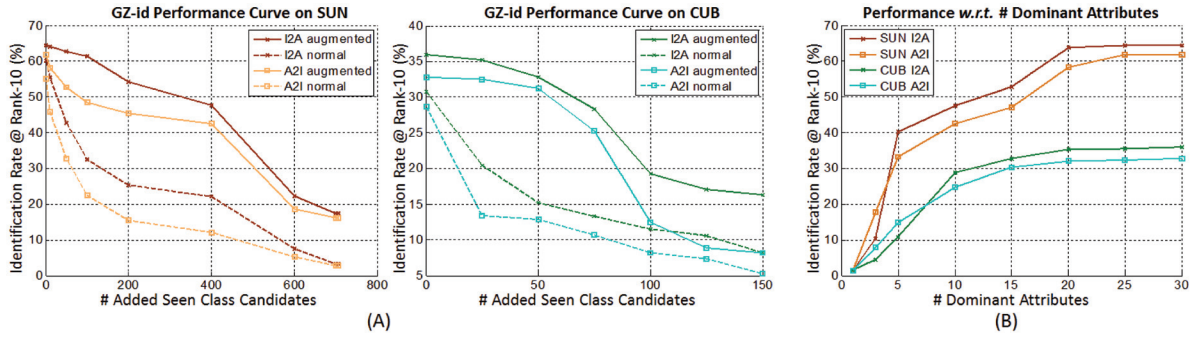

Figure 3: (A) Generalised ZSR (GZSR) performance on SUN and CUB. (B) ZSR performance *w.r.t.* number of selected dominant attributes.

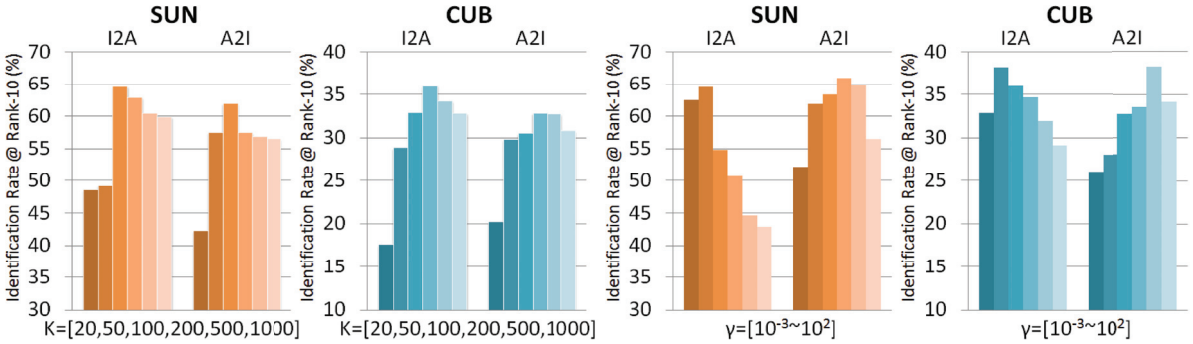

Figure 4: Hyper-parameter Evaluations

lect up to 20 attributes with highest confidence scores as the dominant for each unseen instance. Note that some instances may not have 20 positive attributes. We infer an AA representation for each instance. Since all of the state-of-the-art methods are classification based, we simply regard each unseen instance as a unique class and rank the instances as the original papers. For CCA, NoV, and our method, we use the nearest neighbour criterion to rank the instances as introduced in Eq. 16 and 17.

As shown in Table 1, our method steadily outperforms all of the compared methods at most of the ranks. The exception at rank 1 of ESZSL is due to that some instances can coincidently share the same dominant attributes. The results mainly support the effectiveness of our latent instance-attributes. Unfortunately, such instance-attributes cannot be directly applied to conventional ZSL methods since our representations and projections are jointly learnt. Besides, our results are significantly higher than those of CCA and NoV, which demonstrates the effectiveness of our orthogonal regularisation in LIAD.

Detailed Analysis

In order to further understand the promising results, we discuss the following key questions in detail. The analysis is based on the retrieval performance at rank 10 using up to 20 dominant attributes of each instance.

Generalised ZSR performance Recently, generalised ZSL (Chao et al. 2016) is proposed to investigate whether adding training classes as candidates can degrade the performance. In Fig. 3 (A), we show our generalised ZSR performance on both of the datasets. We also select up to 20 attributes with highest confidence scores as the dominant attributes of each seen instance and encode them into an augmented attribute representation. We then gradually add candidates from seen classes to the gallery and test by the same queries. As shown in Fig. 3 (A), while the performance using normal attributes suffers from severe degradation, using the augmented attributes can significantly improve the tolerance of additional seen candidates. Such results support our assumption that by considering the relations between classes, better generalisation is achieved.

Performance w.r.t. Number of Dominant Attributes An important question is how many dominant attributes are sufficient for the ZSR problem. In Fig. 3 (B), we show hit rates of using from one dominant attribute up to 30 attributes. It can be seen that our method can achieve steady performance after adopting 20 dominant attributes. Using five and ten dominant attributes for each instance on SUN and CUB can still lead to acceptable hit rates.

Hyper-parameter Evaluations There are two main hyper-parameters in our LIAD algorithm. K controls how many latent attributes are sufficient, and γ balances the weight between visual and semantic spaces. In Fig. **??**fig-hyperP, we demonstrate how the two hyper-parameters can affect the

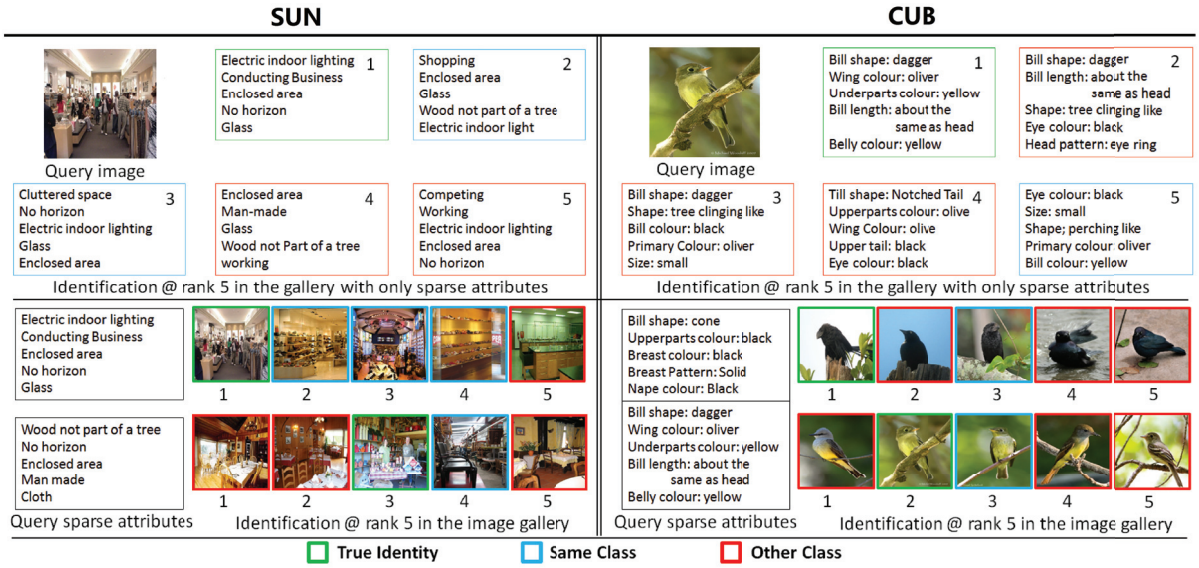

Figure 5: Qualitative Results on SUN and CUB.

Table 3: Comparison with state-of-the-art methods on conventional ZSL settings.

Method	Attributes	SUN		CUB	
DAP (Lampert, Nickisch, and Harmeling 2009)	BA/-	72.0	-	40.3	-
ALE (Akata et al. 2013)	BA/AA	74.	75.8	38.47	41.8
ESZSL (Romera-Paredes and Torr 2015)	BA/AA	82.1	84.6	44.32	48.2
LatEm (Xian et al. 2016)	BA/AA	81.5	83.2	**45.08**	47.6
CCA	BA/AA	73.2	77.2	39.66	42.7
Ours	BA/AA	**82.9**	**85.8**	43.95	**50.2**

ZSR performance. It can be seen that more latent attributes are safer but not necessary to boost the performance. The CUB requires a larger γ than that of the SUN. A larger γ can also benefit the A2I scenario since the gallery tends to be closer to the input dominant attributes.

Performance on ZSL classification Since the state-of-the-art methods are classification-based, we also evaluate whether the inferred attributes are useful for ZSL. Simply, on the I2A scenario, we check whether the top-1 rank and query image are from the same class. To verify the contributions of our augmented attributes (AA), we use conventional binary attributes (BA) as baselines. As shown in Table 3, our results significantly outperform that of compared state-of-the-art methods. Using AA can substantially boost the performance for all of the methods.

Qualitative Results We provide some qualitative results in Fig. 5, which are based on rank 5 using 20 dominant attributes. Due to the length limitation, we can only show top 5 dominant attributes. As can be seen, our method can differentiate very close instances that most of the dominant attributes are shared.

Conclusion

This paper has introduced the ZSR problem. Using the proposed framework, an unseen instance with several dominant attributes can be retrieved in the gallery despite training examples of unseen classes being unavailable. By using word embedding, we have successfully circumvented the requirement of expensive instance-attribute annotations at the training phase. Flexible augmented attributes were proposed to encode either a singular word, multiple words, or an instance described by dominant attributes. Such augmented attributes have shown better generalisation than conventional ones. Furthermore, the LIAD algorithm has been proposed to discover instance-sensitive latent attributes from general class-attributes. On the SUN and CUB benchmarks, our method has demonstrated prominent performance in both I2A and A2I scenarios.

For future work, it is worth to investigate the ZSR performance on conventional instance retrieval tasks. There are two challenges which remain unsolved: 1) circumventing annotation of test instances, and 2) improving the rank-1 ZSR rates by removing misleading shared dominant attributes. Our LIAD can also be applied for conventional ZSL so as to achieve cheap instance attributes from the class-level ones.

References

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2013. Label-embedding for attribute-based classification. In *CVPR*.

Al-Halah, Z.; Tapaswi, M.; and Stiefelhagen, R. 2016. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *CVPR*.

Arandjelovic, R., and Zisserman, A. 2012. Multiple queries for large scale specific object retrieval. In *BMVC*.

Biederman, I. 1987. Recognition-by-components: a theory of human image understanding. *Psychological review* 94(2):115.

Changpinyo, S.; Chao, W.-L.; Gong, B.; and Sha, F. 2016. Synthesized classifiers for zero-shot learning. In *CVPR*.

Chao, W.-L.; Changpinyo, S.; Gong, B.; and Sha, F. 2016. An

empirical study and analysis of generalized zero-shot learning for object recognition in the wild. *arXiv preprint arXiv:1605.04253*.

Elhoseiny, M.; Saleh, B.; and Elgammal, A. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *CVPR*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence* 28(4):594–611.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*.

Haque, A.; Alahi, A.; and Fei-Fei, L. 2016. Recurrent attention models for depth-based person identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jayaraman, D., and Grauman, K. 2014. Zero-shot recognition with unreliable attributes. In *NIPS*.

Jayaraman, D.; Sha, F.; and Grauman, K. 2014. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*.

Kumar, N.; Berg, A. C.; Belhumeur, P. N.; and Nayar, S. K. 2009. Attribute and simile classifiers for face verification. In *ICCV*.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2014. Attribute-based classification for zero-shot visual object categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 36(3):453–465.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*.

Long, Y., and Shao, L. 2017a. Describing unseen classes by exemplars: Zero-shot learning using grouped simile ensemble. In *WACV*.

Long, Y., and Shao, L. 2017b. Learning to recognise unseen classes by a few similes. In *ACMMM*.

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

Long, Y.; Liu, L.; Shen, F.; Shao, L.; and Li, X. 2017b. Zero-shot learning using synthesised unseen visual data with diffusion regularisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Long, Y.; Zhu, F.; Shao, L.; and Han, J. 2017c. Face recognition with a small occluded training set using spatial and statistical pooling. *Information Sciences*.

Long, Y.; Liu, L.; and Shao, L. 2016. Attribute embedding with visual-semantic ambiguity removal for zero-shot learning. In *BMVC*.

Long, Y.; Liu, L.; and Shao, L. 2017. Towards fine-grained open zero-shot learning: Inferring unseen visual features from attributes. In *WACV*.

Long, Y.; Zhu, F.; and Shao, L. 2016. Recognising occluded multi-view actions using local nearest neighbour embedding. *Computer Vision and Image Understanding* 144:36–45.

McCann, S., and Lowe, D. G. 2012. Local naive bayes nearest neighbor for image classification. In *CVPR*.

Mensink, T.; Gavves, E.; and Snoek, C. 2014. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G. S.; and Dean, J. 2014. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*.

Parikh, D., and Grauman, K. 2011. Relative attributes. In *ICCV*.

Patterson, G.; Xu, C.; Su, H.; and Hays, J. 2014. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision* 108(1-2):59–81.

Qiao, R.; Liu, L.; Shen, C.; and van den Hengel, A. 2016. Less is more: Zero-shot learning from online textual documents with noise suppression. In *CVPR*.

Qin, J.; Liu, L.; Shao, L.; Shen, F.; Ni, B.; Chen, J.; and Wang, Y. 2017. Zero-shot action recognition with error-correcting output codes. In *CVPR*.

Rohrbach, M.; Stark, M.; Szarvas, G.; Gurevych, I.; and Schiele, B. 2010. What helps where–and why? semantic relatedness for knowledge transfer. In *CVPR*.

Romera-Paredes, B., and Torr, P. 2015. An embarrassingly simple approach to zero-shot learning. In *ICML*.

Sharmanska, V.; Quadrianto, N.; and Lampert, C. H. 2012. Augmented attribute representations. In *ECCV*.

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

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. 2013. Zero-shot learning through cross-modal transfer. In *NIPS*.

Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*.

Tao, R.; Gavves, E.; Snoek, C. G.; and Smeulders, A. W. 2014. Locality in generic instance search from one example. In *CVPR*.

Tao, R.; Smeulders, A. W.; and Chang, S.-F. 2015. Attributes and categories for generic instance search from one example. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 177–186.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wang, X.; Yang, M.; Cour, T.; Zhu, S.; Yu, K.; and Han, T. X. 2011. Contextual weighting for vocabulary tree based image retrieval. In *ICCV*.

Wen, Z., and Yin, W. 2013. A feasible method for optimization with orthogonality constraints. *Mathematical Programming* 142(1-2):397–434.

Wright, S., and Nocedal, J. 1999. Numerical optimization. *Springer Science* 35:67–68.

Wu, J., and Rehg, J. M. 2008. Where am i: Place instance and category recognition using spatial pact. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.

Xian, Y.; Akata, Z.; Sharma, G.; Nguyen, Q.; Hein, M.; and Schiele, B. 2016. Latent embeddings for zero-shot classification. In *CVPR*.

Yu, X., and Aloimonos, Y. 2010. Attribute-based transfer learning for object categorization with zero/one training example. In *ECCV*.

Zhang, Z., and Saligrama, V. 2015. Zero-shot learning via semantic similarity embedding. In *ICCV*.