Attributes2Classname: A discriminative model for attribute-based unsupervised zero-shot learning

Berkan Demirel[1,3], Ramazan Gokberk Cinbis[2], Nazli Ikizler-Cinbis[3]

[1]HAVELSAN Inc., [2]Bilkent University, [3]Hacettepe University

bdemirel@havelsan.com.tr, gcinbis@cs.bilkent.edu.tr, nazli@cs.hacettepe.edu.tr

Abstract

We propose a novel approach for unsupervised zero-shot learning (ZSL) of classes based on their names. Most existing unsupervised ZSL methods aim to learn a model for directly comparing image features and class names. However, this proves to be a difficult task due to dominance of non-visual semantics in underlying vector-space embeddings of class names. To address this issue, we discriminatively learn a word representation such that the similarities between class and combination of attribute names fall in line with the visual similarity. Contrary to the traditional zero-shot learning approaches that are built upon attribute presence, our approach bypasses the laborious attribute-class relation annotations for unseen classes. In addition, our proposed approach renders text-only training possible, hence, the training can be augmented without the need to collect additional image data. The experimental results show that our method yields state-of-the-art results for unsupervised ZSL in three benchmark datasets.

1. Introduction

Zero-shot learning (ZSL) enables identification of classes that are not seen before by means of transferring knowledge from seen classes to unseen classes. This knowledge transfer is usually done via utilizing prior information from various auxiliary sources, such as attributes (*e.g.* [20, 12, 27, 5, 35, 6, 4]), class hierarchies (*e.g.* [27]), vector-space embeddings of class names (*e.g.* [35, 4, 6]) and textual descriptions of classes (*e.g.* [22, 10]). Among these, attributes stand out as an excellent source of prior information: (i) thanks to their visual distinctiveness, it is possible to build highly accurate visual recognition models of attributes; (ii) being linguistically descriptive, attributes can naturally be used to encode classes in terms of their visual appearances, functional affordances or other human-understandable aspects.

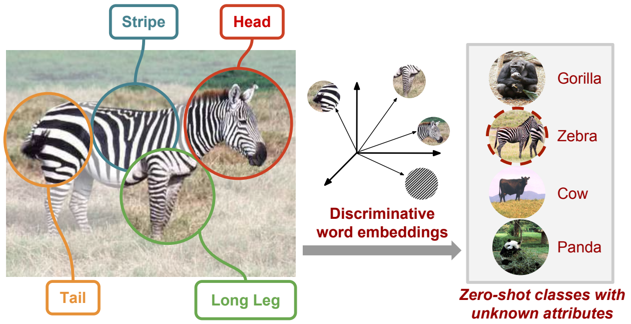
Stripe
Head
Tail
Long Leg
Discriminative word embeddings
Gorilla
Zebra
Cow
Panda
Zero-shot classes with unknown attributes

Figure 1: We propose a zero-shot recognition model based on attribute and class names. Unlike most other attribute-based methods, our approach avoids the laborious attribute-class relations at test time, by discriminatively learning a word-embedding space for predicting the unseen class name, based on combinations of attribute names.

Almost all attribute-based ZSL works, however, have an important disadvantage: attribute-class relations need to be precisely annotated not only for the seen (training) classes, but also for the unseen (zero-shot) classes (*e.g.* [12, 20, 27, 5]). This usually involves collecting fine-grained information about attributes and classes, which is a time-consuming and error-prone task limiting the scalability of the approaches to a great extent.

Several recent studies explore other sources of prior information to alleviate the need of collecting annotations at test time. These approaches rely on readily available sources like word embeddings and/or semantic class hierarchies, hence, do not require dedicated annotation efforts. We simply refer to these as *unsupervised ZSL*. Such approaches, however, exclude attributes at the cost of exhibiting a lower recognition performance [4].

Towards combining the practical merit of unsupervised ZSL with the recognition power of attribute-based methods, we propose an attribute-based unsupervised ZSL approach. The main idea is to discriminatively learn a vector-space representation of words in which the combination of attributes relating to a class and the corresponding class name are mapped to nearby points. In this manner, the model would map distinctive attributes in images to a semantic word vector space, using which we can predict unseen classes solely based on their names. This idea is illustrated in Figure 1.

Our use of vector space word embeddings differs significantly from the way they are used in existing unsupervised ZSL methods: existing approaches (*e.g.* [35, 4]) aim to build a comparison function directly between image features and class names. However, learning such a comparison function is difficult since word embeddings are likely to be dominated by non-visual semantics, due to lack of visual descriptions in the large-scale text corpora that is used in the estimation of the embedding vectors. Therefore, the resulting zero-shot models also tend to be dominated by non-visual cues, which can degrade the zero-shot recognition accuracy. To address this issue, we propose to use the names of visual attributes as an intermediate layer that connects the image features and the class names in an unsupervised way for the unseen classes.

An additional interesting aspect of our approach is the capability of *text-only training*. Given pre-trained attribute models, the proposed ZSL model can be trained based on textual attribute-class associations, without the need for explicit image data even for training classes. This gives an extreme flexibility for scalability: the training set can be easily extended by enumerating class-attribute relationships, without the need for collecting accompanying image data. The resulting ZSL model can then be used for recognition of zero-shot classes for which no prior attribute information or visual training example is available.

We provide an extensive experimental evaluation on two ZSL object recognition and one ZSL action recognition benchmark datasets. The results indicate that the proposed method yields state-of-the-art unsupervised zero-shot recognition performance both for object and cross-domain action recognition. Our unsupervised ZSL model also provides competitive performance compared to the state-of-the-art supervised ZSL methods. In addition, we experimentally demonstrate the success of our approach in the case of text-only training. Finally, the qualitative results suggest that the non-linear transformation of the proposed approach improves visual semantics of word embeddings, which can facilitate further research.

To sum up, our main contributions are as follows: (i) we propose a novel method for discriminatively learning a word vector space representation for relating class and attribute combinations purely based on their names. (ii) We show that the learned non-linear transformation improves the visual semantics of word vectors. (iii) Our method achieves the state-of-the-art performance among unsupervised ZSL approaches and (iv) we show that by augmenting the training dataset by additional class names and their attribute predicate matrices but no visual examples, a boost in performance can be achieved.

2. Related work

Initial attempts towards zero-shot classification were supervised, in the sense that they require explicit attribute annotations of the test classes (*e.g.* [21, 20, 5, 27, 9, 16, 29, 36, 38, 39]). Lampert *et al.* [21, 20] are among the first to use attributes in this setting. They propose direct (DAP) and indirect attribute prediction (IAP) where attribute and class relations are provided explicitly. Al-Halah *et al.* [5] introduce hierarchy and apply attribute label propagation on object classes, to utilize attributes at different abstraction levels. Rohrbach *et al.* [27] propose a similar hierarchical method, but they use only class taxonomies. Deng *et al.* [9] introduce Hierarchy and Exclusion (HEX) graphs as a standalone layer to be used on top of any-feedforward architecture for classification. Jayaraman and Grauman [16] propose a random forest approach to handle error tendencies of attributes. Romera *et al.* [29] develop two linear layered network to handle relations between classes, attributes and features. Zhang and Saligrama [36] propose a method to use semantic similarity embedding where target classes are represented with histograms of the source classes.

An important limitation of the aforementioned methods is their dependency on the attribute signatures of the test classes. To apply these approaches to additional unseen classes, the attribute signatures of those new classes need to be provided explicitly. Our method alleviates this need by learning a word representation that allows zero-shot classification by comparing class names and attribute combinations, with no explicit prior information about attribute relations of unseen classes.

Recently, unsupervised ZSL methods are gaining more attention, due to their increased scalability. Instead of using class-attribute relations at test time, various auxiliary sources of side information, such as textual information [22, 10] or word embeddings [3, 4, 25, 14, 6, 8] are explored in such methods. Ba *et al.* [22] propose to combine MLP and CNN networks handling text based information acquired from Wikipedia articles and visual information of images, respectively. Another interesting direction is explored by Elhoseiny *et al.* [10], where the classifiers are built directly on textual corpus that is accompanied with images.

Distributional word representations, or word embeddings, [23, 24, 26] are becoming increasingly popular [3, 4, 25, 14], due to the powerful vector-space representations where the distances can be meaningfully utilized. Akata *et al.* [3] propose attribute label embedding (ALE) method that uses textual data as side information in the WSABIE [34] formulation. Akata *et al.* [4] improve ALE by using embedding vectors that were obtained from large-scale text corpora. Frome *et al.* [14] propose a similar model where a pre-trained CNN model is fine-tuned in an end-to-end way to relate images with semantic class embeddings. Norouzi *et al.* [25] proposes to use convex combinations of fixed class name embeddings, weighted by class posterior probabilities given by a pre-trained CNN model, to map images to the class name embedding space. In the recent approach of Akata *et al.* [2] language representations are utilized jointly with the stronger supervision given by visual part annotations. Xian *et al.* [35] use multiple visual embedding spaces to encode different visual characteristics of object classes. Jain *et al.* [15] and Kordumova *et al.* [18] leverage pre-trained object classifiers, and, action-object similarities given by class embeddings to assign action labels to unseen videos.

The work closest to ours is Al-Halah *et al.* [6], which proposes an approach for using visual attributes in the unsupervised ZSL setting. In their approach, a model is learned to predict whether an individual attribute is related to a class name or not. For this purpose, they learn a separate bilinear compatibility function for each group of attributes, where similar attributes are grouped together to improve the performance. For unsupervised ZSL, this approach first estimates the association of attributes with the test class, and then employs an attribute-based ZSL method using the estimated class-attribute relations. Our approach differs in two major ways. First, instead of comparing classes with individual attribute names, we model the relationship between class names and combinations of attribute names. Second, as opposed to handling class-attribute relation estimation and zero-shot classification as two separate problems, we discriminatively train our attribute based ZSL model in an end-to-end manner.

3. Method

In this section, we present the details of our approach. First, we explain our zero-shot learning model. Then, we describe how to train our ZSL model using discriminative *image-based training* and *predicate-based training* formulations. Finally, we briefly discuss our *text-only training* strategy for incorporating additional classes during training.

3.1. Zero-shot learning model

We define our ZSL model compatibility function $f(x,y) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{R}$ that measures the relevance of label $y \in \mathcal{Y}$ for a given image $x \in \mathcal{X}$. Using this function, a test image x can be classified simply by choosing the class maximizing the compatibility score: $\arg\max_y f(x,y)$.

In order to enable zero-shot learning of classes based on class names only, we assume that an initial d_0-dimensional vector space embedding $\varphi_y \in \mathcal{R}^{d_0}$ is available for each class y. These initial class name embeddings are obtained using general purpose corpora, due to lack of a large-scale text corpus dedicated for visual descriptions of objects. The representations obtained by the class embeddings, hence, are typically dominated by non-visual semantics. For instance, according to the GloVe vectors, the similarity between *wolf* and *bear* (both wild animals) is higher that the similarity between *wolf* and *dog*, though the latter pair is visually much more similar to each other.

These observations suggest that learning a compatibility function directly between the image features and class embeddings may not be easy due to non-visual components of word embeddings. To address this issue, we propose to leverage attributes, which are appealing for the dual representation they provide: each attribute corresponds to (i) a visual cue in the image domain, and, (ii) a named entity in the language domain, whose similarity with class names can be estimated using word embeddings. We define a function $\Phi(x) : \mathcal{X} \rightarrow \mathcal{R}^d$ for embedding each image based on the attribute combination associated with it:

$$\Phi(x) = \frac{1}{\sum_a p(a|x)} \sum_a p(a|x) T(\varphi_a) \tag{1}$$

where $p(a|x)$ is the posterior probability of attribute a[1], given by a pre-trained binary attribute classifier, φ_a is the initial embedding vector of attribute a, and $T : \mathcal{R}^{d_0} \rightarrow \mathcal{R}^d$ is the transformation that we aim to learn. Similarly, we define our class embedding function $\phi(y) : \mathcal{Y} \rightarrow \mathcal{R}^d$ as the transformation of the initial class name embeddings φ_y: $\phi(y) = T(\varphi_y)$.

The purpose of the function T is to transform the initial word embeddings of attributes and classes such that each image, and its corresponding class are represented by nearby points in the d-dimensional vector embedding space. Consequently, we can define $f(x,y)$ as a similarity measure between the image and class embeddings. In our approach, we opt for the cosine-similarity:

$$f(x,y) = \frac{\Phi(x)^{\mathrm{T}} \phi(y)}{\|\Phi(x)\| \|\phi(y)\|} \tag{2}$$

We emphasize that our approach requires only the name of an unseen class at test time, as the compatibility function relies solely on the learned attribute and class name embeddings, rather than attribute-class relations.

Figure 2 illustrates our zero-shot classification approach. Given an image, we first apply the attribute predictors and compute a weighted average of the attribute name embeddings. The class assignment is done by comparing the

[1] The normalization in the denominator aims to make the embeddings comparable across images with varying number of observed attributes.

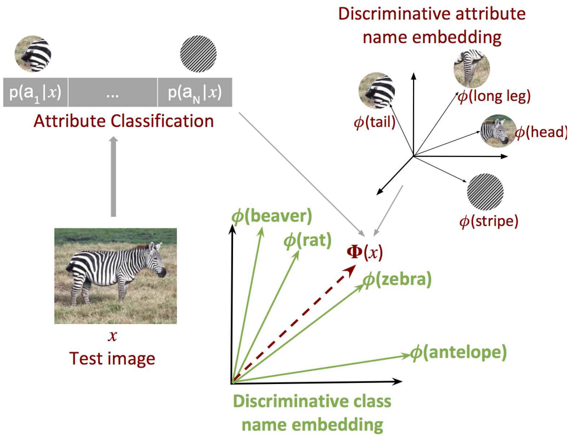

Figure 2: Illustration of our unsupervised zero-shot recognition model. Prediction depends on the similarity between discriminatively learned representations of attribute combinations and class names. (Best viewed in color.)

resulting embedding of attribute combination with that of each (unseen) class name. The image is then assigned to the class with the highest cosine similarity.

As defined above, the embeddings of attribute combinations and class names are functions of the shared transformation $T(\varphi)$.[2] In our experiments, we define $T(\varphi)$ as a two-layer feed-forward neural network. In the following sections, we describe techniques for discriminatively learning this transformation network.

3.2. Image-based training (IBT)

In image-based training, we assume that there exists a supervised training set S of N examples. Each example forms an image and class label pair. By definition, no example in S belongs to one of the zero-shot test classes. Our goal is to discriminatively learn the function $f(x,y)$ such that for each training example i, the compatibility score of the correct class $y = y_i$ is higher than any other class y_j, by a margin of $\Delta(y_i, y_j)$. More formally, the training constraint for the i-th training example is given by

$$f(x_i, y_i) \geq f(x_i, y_j) + \Delta(y_i, y_j), \quad \forall y_j \neq y_i \qquad (3)$$

The margin function Δ indicates a non-negative pairwise discrepancy value for each pair of the training classes.

As explained in the previous section, $f(x,y)$ is a function of the transformation network $T(\varphi)$. Let θ be the vector of all parameters in the transformation network. Inspired from the structural SVMs [33, 28], we formalize our approach as a constrained optimization problem:

$$\min_{\theta,\xi} \lambda\|\theta\| + \sum_{i=1}^{N} \sum_{y_j \neq y_i} \xi_{ij}$$

$$f(x_i, y_i) \geq f(x_i, y_j) + \Delta(y_i, y_j) - \xi_{ij} \quad \forall y_j \neq y_i, \forall i \qquad (4)$$

where ξ is a vector of slack variables for soft-penalizing unsatisfied similarity constraints, and λ is the regularization weight. To avoid optimization over non-linear constraints, we can equivalently express this problem as an unconstrained optimization problem:

$$\min_{\theta} \lambda\|\theta\|_2^2 + \sum_{i=1}^{N} \sum_{y_j \neq y_i} \max\left(0, f(x_i, y_j) - f(x_i, y_i) + \Delta(y_i, y_j)\right) \qquad (5)$$

Using this formulation, the transformation $T(\varphi)$ is learned in an discriminative and end-to-end manner, by ensuring that the correct class score is higher than the incorrect ones, for each image.

We empirically observe that cross-validating the number of iterations provides an effective regularization strategy, therefore, we fix $\lambda = 0$. We use average Hamming distance between the attribute indicator vectors, which denote the list of attributes associated with each class, to compute Δ values. This is the only point where we utilize the class-attribute predicate matrix in our image-based training approach. In the absence of a predicate matrix, other types of Δ functions, like word embedding similarities, may be explored, which we leave for future work. Other implementation details are provided in Section 4.

3.3. Predicate-based training (PBT)

In this section, we propose an alternative training approach, which we call predicate-based training. In this approach, the goal is to learn the ZSL model solely based on the predicate matrix, which denotes the class-attribute relations. While image-based training is defined in terms of image-class similarities, we formulate predicate-based training in terms of class-class similarities, without directly using any visual examples during training.

The predicate matrix consists of per-class indicator vectors, where each element is one if the corresponding attribute is associated with the class, and zero, otherwise. We denote the indicator vector for class y by π_y. Then, similar to image embedding function $\Phi(x)$, we define a *predicate-embedding* function $\Psi(\pi)$:

$$\Psi(\pi) = \frac{1}{\sum_a \pi(a)} \sum_a \pi(a) T(\varphi_a). \qquad (6)$$

This embedding function is obtained by replacing posterior probabilities in Eq. (1) by binary attribute-class relations. Then, we define a new compatibility function $g(\pi, y)$, as the cosine similarity between the vector $\Psi(\pi)$ and vector

[2] In principle, one can separately define a $T(\varphi)$ for attribute names, and, another one for class names. We have explored this empirically, but did not observe a consistent improvement. Therefore, for the sake of simplicity, we use a shared transformation network in our experiments.

$\phi(y)$. This function is basically similar to Eq. (2), where the image embedding $\Phi(x)$ is replaced by the attribute indicator embedding $\Psi(\pi)$.

Finally, we define the learning problem as optimizing the function $g(x, y)$ such that for each class, the compatibility score for its ideal set of attributes π_y is higher than the attribute combination $\pi_{y'}$ of another class y', by a margin of $\Delta(y, y')$. This constraint aims to ensure that the similarity between the name embedding of a set of attributes and the embedding of a class name reliably indicates the visual similarity indicated by the predicate matrix.

This definition leads us to an unconstrained optimization problem analogous to Eq. (5):

$$\min_\theta \lambda\|\theta\|_2^2 + \sum_{y=1}^{K} \sum_{y' \neq y_i} \max\left(0, g(\pi_{y'}, y_i) - g(\pi_{y_i}, y_i) + \Delta(y_i, y')\right) \tag{7}$$

where K indicates the number of training classes in the predicate matrix. As in image-based training, we define $\Delta(y, y')$ as the average Hamming distance between π_y and $\pi_{y'}$, and use $\lambda = 0$.

Figure 3 illustrates the predicate-based training approach. As shown in this figure, the main idea is to project the φ word representations into a new space, where the similarity between a class and an attribute combination in terms of their name vectors is indicative of their visual similarity. At test time, we use the learned transformation network in zero-shot classification via the compatibility function $f(x, y)$ in Eq. (2). This compatibility function uses only attribute classifier outputs and the transformed word embeddings.

3.4. Text-only training

Predicate-based training, as explained in the previous section, is completely based on a class-attribute predicate matrix for the training classes, and training images are used only for pre-training attribute classifiers that will be used at test time. In contrast, image-based training, directly learns the ZSL model based on attribute classification probabilities in training images, therefore in principle, we expect image-based training to perform better. This is, in fact, verified in our experimental results: while predicate-based training shows competitive accuracy, we obtain our state-of-the-art results using image-based training.

Despite the relatively lower performance of predicate-based training, it has one interesting property: we can expand the training set by simply adding textual information for additional novel classes into the predicate matrix. This allows improving the ZSL model by using classes with no visual examples. We call incorporation of additional training classes in this manner as *text-based training*. In Section 4, we empirically show that it is possible to improve the predicate-based training using text-based training.

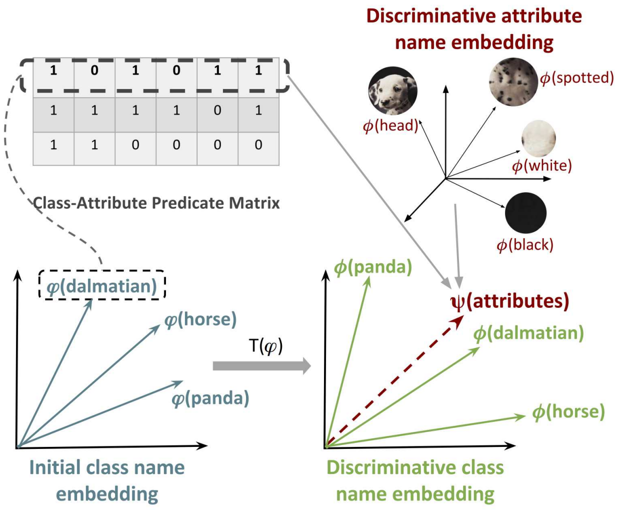

Figure 3: Illustration of our predicate-based training approach, which uses only the predicate matrix of class and attribute relations as the source of supervision. The goal is to represent class and attribute combinations, based on their names, in a space where each class is closest to its ideal attribute combination.

4. Experiments

To evaluate the effectiveness of the proposed approach, we consider two different ZSL applications: zero-shot object classification and zero-shot action recognition.

4.1. Zero Shot Object Classification

In this part, we explain our zero-shot object classification experiments on two common datasets namely AwA [20], aPaY [13]. AwA dataset [20] contains 30,475 images of 50 different animal classes. 85 per-class attribute labels are provided in the dataset. In the predefined split for zero-shot learning, 40 animal classes are marked for training and 10 classes for testing. aPaY dataset [13] is formed of images obtained from two different sources. aPascal (aP) part of this dataset is obtained from PASCAL VOC 2008 [11]. This part contains 12,695 images of 20 different classes. The second part, aYahoo (aY), is collected using Yahoo search engine and contains 2,644 images of 12 object classes completely different from aPascal classes. Images are annotated with 64 binary per-image attribute labels. In zero-shot learning settings on this dataset, aPascal part is used for training and aYahoo part is used for testing. We follow the same experimental setup as in [5] and only use training split of aPascal part to learn attribute classifiers.

Attribute Classifiers. We use CNN-M2K features [5] to encode images and train attribute classifiers. We resize each image to 256x256 and then subtract the mean image. Data

Table 1: Zero-shot classification performance of proposed predicate-based (PBT) and image-based (IBT) methods on AwA and aPaY datasets. We report normalized accuracy.

Method	AwA	aPaY
Baseline	10.2	16.0
PBT	60.7	29.4
IBT	**69.9**	**38.2**

augmentation is applied via using five different crops and their flipped versions. Outputs of fc7 layer are used, resulting in 2,048 dimensional feature vectors. Following [13], we obtain the attribute classifiers by training ℓ_2-regularized squared-hinge-loss linear SVMs. Parameter selection is done using 10-fold cross validation over the training set and Platt scaling is applied to map the attribute prediction scores to posterior probabilities. For image-based training, cross-validation outputs are used as the classification scores in training images.

Word Embeddings. For each class and attribute name, we generate a 300-dimensional word embedding vector using GloVe [26] based on Common Crawl Data[3]. These word vectors are publicly available[4]. For those names that consist of multiple words, we use the average of the word vectors.

Word Representation Learning. We define the transformation function as a two layer feed-forward network. We use 2-fold cross-validation over the training set to select number of hidden units and number of iterations. *tanh* function is used as the activation function in the first hidden layer and *sigmoid* function is used in the second hidden layer. Adam [17] is used for stochastic optimization, and learning rate value is set to 1e-4. Implementation is done using TensorFlow [1].[5]

Results. In our experiments, we first evaluate the performance of attribute classifiers, since this is likely to have a significant influence on zero-shot classification. The attribute classifiers yield 80.56% mean AUC on the AwA dataset, 84.91% mean AUC on the aPaY dataset. These results suggest that our attribute classifiers are relatively accurate, if not perfect. Further improvements in attribute classification are likely to have a positive impact on the final ZSL performance.

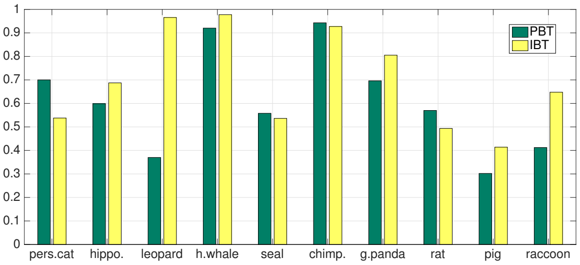

Figure 4: Class-wise prediction accuracies on AwA Dataset.

Table 1 presents the experimental results for our approach. In this table, *baseline* represents the case where the transformation $T(\varphi)$ is defined as an identity mapping. PBT (predicate-based training) represents our proposed approach that learns a transformation using the attribute predicate matrix, whereas IBT (image-based training) represents learning transformation using training images. The results in Table 1 shows the importance and success of our learning formulations, compared to the baseline. In addition, we observe that image-based training outperforms predicate-based training on average, which is in accordance with our expectations. Class-wise accuracy comparison of PBT and IBT methods is given in Figure 4. We observe that some of the classes respond particularly well to the image-based training.

Table 2 presents a comparison of our results against a number of supervised and unsupervised ZSL methods. In this table, the supervision corresponds to the information needed during test time for zero-shot learning: the supervised methods require additional data about the unseen classes such as attribute-class predicate matrices, whereas unsupervised methods do not require any explicit inputs about test classes. Hence, supervised methods have a very major advantage in this comparison, as they employ external attribute signatures of test classes. In contrast, unsupervised methods carry out zero-shot classification among the test classes without using data additional to the training set. Finally, we note that, we exclude ZSL methods that operate on low-level visual image features, as their results are not directly comparable. Instead, for the sake of fair comparison, we only compare to those methods that use similar convolutional neural network based image representations.

From Table 2 we see that on AwA and aPaY datasets, our unsupervised ZSL method yields state-of-the-art classification performance compared to other unsupervised ZSL methods. In addition, our method performs on par with some of the supervised ZSL methods.

4.2. Zero Shot Action Recognition

For zero-shot action recognition, we evaluate our approach on UCF-Sports Action Recognition Dataset [30]. The dataset is formed of videos from various sport actions which are featured from television channels such as the BBC and ESPN, and contains a total of 150 videos of 10 different sport action classes.

[3] commoncrawl.org/the-data/

[4] nlp.stanford.edu/projects/glove/

[5] github.com/berkandemirel/attributes2classname

Table 2: Comparison to state-of-the-art ZSL methods (unsupervised and supervised).

Test supervision	Method	AwA	aPaY
unsupervised	DeViSE[14]	44.5	25.5
	ConSE[25]	46.1	22.0
	Text2Visual[10, 7]	55.3	30.2
	SynC[8]	57.5	-
	ALE[4]	58.8	33.3
	LatEm[35]	62.9	-
	CAAP[6]	67.5	37.0
	Our method	**69.9**	**38.2**
supervised	DAP[20]	54.0	28.5
	ENS[27]	57.4	31.7
	HAT[5]	63.1	38.3
	ALE-attr[4]	66.7	-
	SSE-INT[36]	71.5	44.2
	SSE-ReLU[36]	76.3	46.2
	SynC-attr[8]	76.3	-
	SDL[38]	79.1	50.4
	JFA[37]	81.0	52.0

Word Embeddings. Following [15], we utilize 500-dimensional word embedding vectors generated with the skip-gram model of word2vec [23] learned over YFCC100M [32] dataset. YFCC100M dataset contains metadata tags of about 100M Flickr images and the word vectors obtained from YFCC100M are publicly available[6].

Object Classifiers. Since there is no explicit definition of attributes for actions, the object cues can be leveraged instead of attributes, as suggested by [15]. To this end, we obtain predicate matrices from the textual data by measuring the cosine similarity between actions and object classification scores. We operate on the object classification responses made available by [15][6]. These are obtained by AlexNet[19], where every 10th frame is sampled for each video and each sampled frame is represented with the total of 15,293 ImageNet object categories. Average pooling is applied afterwards, so that each video is represented with 15,293 dimensional vectors. To have a fair comparison, we also apply the sparsification step of [15] using the same parameters. This sparsification is done for eliminating noisy object classification responses.

Word Representation Learning. Model learning settings are the same with those of ZSL object classification experiments, with the exception that only image-based loss is used, because predicate matrices are not available during training. Since we do not have any training data for target datasets, we train our transformation function with a different dataset (*i.e.* UCF-101 [31]). To avoid any overlap between datasets, we exclude the common action classes from the training set for an accurate zero-shot setting. Some of such common classes that are excluded from training are *Diving* and *Horse Riding*.

Table 3: Zero-shot action recognition accuracies.

Method	UCF-Sport
DAP[20]	11.7
objects2action[15]	26.4
Our method	**28.3**

Table 4: Zero-shot learning using external training class names and their predicate matrices. These EXT classes consist of class names outside AwA dataset and do not include image data. The method is trained only on class names and their predicate matrices. We report normalized accuracy.

Method	Train Classes	Accuracy
PBT	EXT	44.0
PBT	AWA	60.7
PBT	AWA+EXT	63.0

Results. We compare our approach with Objects2Action [15] and DAP [20] methods. The normalized accuracy results are shown in Table 3. From these results we see that our approach for relating action names and object cues in the transformed word vector space yields promising results in UCF-Sport dataset. These results show that our embedding transformation function carries substantial semantic information not only between training and test sets, but also across datasets.

4.3. Training on Textual Data

As stated before, one of the interesting aspects of our formulation is the ability to train over only textual data (*i.e.* names of attributes, objects and classes), without having any visual examples of training classes. In this case, using our model, we can use the pre-trained attribute classifiers, together with the learned semantic word vector representation and predict the class of a newly seen example.

To demonstrate the effect, we select 20 classes outside the AwA dataset from Wikipedia Animal List[7], and build an attribute-class predicate matrix. We then learn the corresponding semantic vector space using only these classes that have no image data. The results are shown in Table 4. Note that, here, we only train the PBT model, because IBT is based on image data. Training our model using only additional textual class names and their corresponding attribute predicate matrices gives an impressive accuracy of 44.0%. Moreover, when we augment the AwA train set with these

[6] staff.fnwi.uva.nl/m.jain/projects/Objects2action.html

[7] en.wikipedia.org/wiki/List_of_animal_names

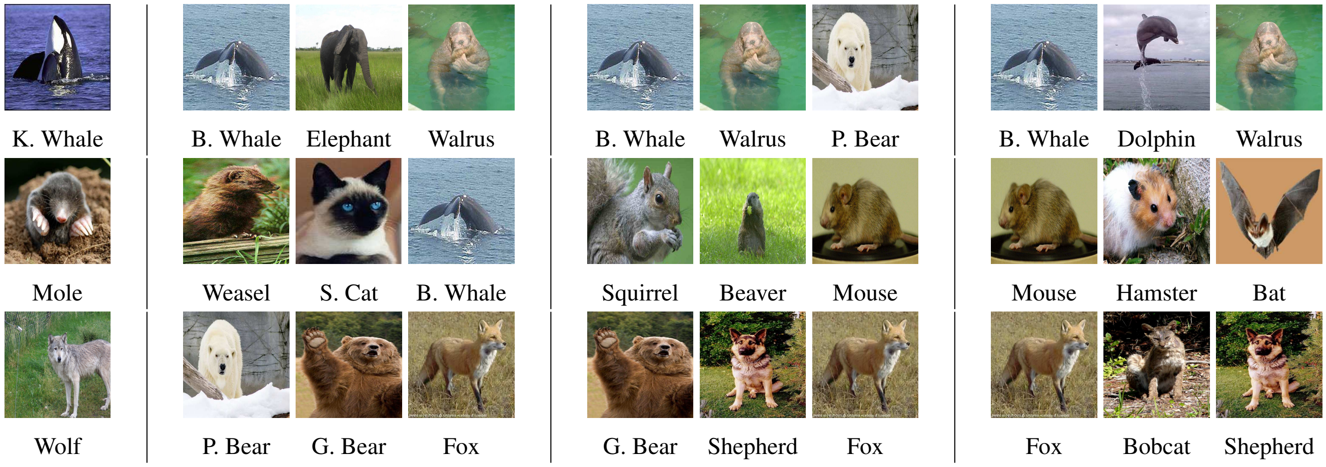

Figure 5: Top-3 most similar classes for some example classes from the AwA dataset. The similarities of the class word vectors are measured by cosine similarity. The images shown depict class representatives. From left-to-right, the columns show the query class (first column), and the most similar classes according to raw word embeddings (second column), those using the transformation learned by PBT (third column), and those using the transformation learned by IBT (fourth column), respectively.

additional class names and their predicate matrix, the accuracy improves from 60.7% to 63.0%. These results suggest that the performance of the proposed model can be improved by just enumerating additional class names and their corresponding attribute lists, without necessarily collecting additional image data.

4.4. Visual Similarities of Word Vectors

One of the favorable aspects of our method is that it can lead to visually more consistent word embeddings of visual entities. To demonstrate this, Figure 5 shows the similarities across the classes according to the original and transformed word embeddings in the AwA dataset. In the first row, we see that while one of the most similar classes to the *killer whale* is *elephant* using the original embeddings, this changes to the *dolphin* class after using the transformation learned by IBT. We observe similar improvements for other classes, such as *mole* (second row) and *wolf* (third row), for which the word embeddings transformed by PBT or IBT training lead to visually more sensible word similarities.

4.5. Randomly Sampled Vectors

To quantify the importance of initial word embeddings, we evaluate our approach on the AwA dataset by using vectors sampled from a uniform distribution, instead of pre-trained GloVe vectors. In this case, PBT yields 28.6%, and IBT yields 13.6% top-1 classification accuracy, which are significantly lower than our actual results (PBT 69.9% and IBT 60.7%). This observation highlights the importance of leveraging prior knowledge derived from unsupervised text corpora through pre-trained word embeddings.

5. Conclusion

An important limitation of the existing attribute-based methods for zero-shot learning is their dependency on the attribute signatures of the unseen classes. To eliminate this dependency, in this work, we leverage attributes as an intermediate representation, in an unsupervised way for the unseen classes. To this end, we learn a discriminative word representation such that the similarities between class and attribute names follow the visual similarity, and use this learned representation to transfer knowledge from seen to unseen classes. Our proposed zero-shot learning method is easily scalable to work with any unseen class without requiring manually defined attribute-class annotations or any type of auxiliary data.

Experimental results on several benchmark datasets demonstrate the efficiency of our approach, establishing the state-of-the-art among the unsupervised zero-shot learning methods. The qualitative results show that the non-linear transformation using the proposed approach improves distributed word vectors in terms of visual semantics. In addition, we show that by adding just text-based class names and their attribute signatures, the training set can be easily extended, which can further boost the performance.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous systems, 2015.

[2] Z. Akata, M. Malinowski, M. Fritz, and B. Schiele. Multicue zero-shot learning with strong supervision. In *Proc.*

IEEE Conf. Comput. Vis. Pattern Recog., pages 59–68, 2016.

[3] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 819–826, 2013.

[4] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2927–2936, 2015.

[5] Z. Al-Halah and R. Stiefelhagen. How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes. In *IEEE Winter Conf. on Applications of Computer Vision*, pages 837–843. IEEE, 2015.

[6] Z. Al-Halah, M. Tapaswi, and R. Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5975–5984, 2016.

[7] L. Bo and C. Sminchisescu. Twin gaussian processes for structured prediction. *Int. J. on Computer Vision*, 87(1-2):28–52, 2010.

[8] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5327–5336, 2016.

[9] 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 *Proc. European Conf. on Computer Vision*, pages 48–64. Springer, 2014.

[10] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 2584–2591, 2013.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 Results.

[12] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009.

[13] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1778–1785. IEEE, 2009.

[14] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 2121–2129, 2013.

[15] M. Jain, J. C. van Gemert, T. Mensink, and C. G. Snoek. Objects2action: Classifying and localizing actions without any video example. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 4588–4596, 2015.

[16] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 3464–3472, 2014.

[17] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. Int. Conf. Learn. Represent.*, 2015.

[18] S. Kordumova, T. Mensink, and C. G. Snoek. Pooling objects for recognizing scenes without examples. In *Proc. ACM Int. Conf. Multimedia Retrieval*, pages 143–150. ACM, 2016.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 1097–1105, 2012.

[20] C. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(3):453–465, March 2014.

[21] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 951–958. IEEE, 2009.

[22] J. Lei Ba, K. Swersky, S. Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 4247–4255, 2015.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 3111–3119, 2013.

[24] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751, 2013.

[25] 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 *Proc. Int. Conf. Learn. Represent.*, 2014.

[26] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. *Proc. of the Empiricial Methods in Natural Language Processing*, 12:1532–1543, 2014.

[27] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1641–1648, 2011.

[28] B. T. C. G. D. Roller. Max-margin markov networks. *Proc. Adv. Neural Inf. Process. Syst.*, 16:25, 2004.

[29] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *Proc. Int. Conf. Mach. Learn.*, pages 2152–2161, 2015.

[30] K. Soomro and A. R. Zamir. Action recognition in realistic sports videos. In *Computer Vision in Sports*, pages 181–208. Springer, 2014.

[31] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human action classes from videos in the wild. Technical Report CRCV-TR-12-01, University of Central Florida, November 2012.

[32] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[33] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6(Sep):1453–1484, 2005.

[34] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.

[35] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 69–77, 2016.

[36] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 4166–4174, 2015.

[37] Z. Zhang and V. Saligrama. Learning joint feature adaptation for zero-shot recognition. *arXiv preprint arXiv:1611.07593*, 2016.

[38] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6034–6042, 2016.

[39] Z. Zhang and V. Saligrama. Zero-shot recognition via structured prediction. In *Proc. European Conf. on Computer Vision*, pages 533–548. Springer, 2016.