A Large-scale Attribute Dataset for Zero-shot Learning

Bo Zhao[1], Yanwei Fu[2], Rui Liang[3], Jiahong Wu[3], Yonggang Wang[3], Yizhou Wang[1]

[1] National Engineering Laboratory for Video Technology, Key Laboratory of Machine Perception (MoE),
Cooperative Medianet Innovation Center, Shanghai,
School of EECS, Peking University, Beijing, 100871, China

[2] School of Data Science, Fudan University, [3] Sinovation Ventures

bozhao, yizhou.wang@pku.edu.cn, yanweifu@fudan.edu.cn, liangrui, wujiahong, wangyonggang@chuangxin.com

Abstract

Zero-Shot Learning (ZSL) has attracted huge research attention over the past few years; it aims to learn the new concepts that have never been seen before. Each concept (class) is embedded in two or more modalities, e.g., the image features and semantic embeddings. Attributes are introduced as the intermediate semantic representation to realize the knowledge transfer from seen to unseen classes. Previous ZSL algorithms are tested on several benchmark datasets, which are defective in terms of the image distribution and attribute diversity. In addition, we argue that the "co-occurrence bias problem" of existing datasets, which is caused by the biased co-occurrence of objects, significantly hinders models from correctly learning the concept. To overcome these problems, we propose a Large-scale Attribute Dataset (LAD) with 78,017 images of 230 classes. 359 attributes of visual, semantic and subjective properties are defined and annotated in instance-level. Seven state-of-the-art ZSL algorithms are tested on this new dataset. The experimental results reveal the challenge of implementing ZSL on our dataset. Based on the proposed dataset, Zero-shot Learning Competition of AI Challenger ($>$ 110 teams attended) has been organized for promoting ZSL research.

1. Introduction

Humans can distinguish more than 30,000 basic level concepts and many more subordinate ones [3], while existing deep neural networks [33, 34, 12] can only classify thousands of objects. It is expensive to collect the labelled data sufficiently to train deep neural networks for all classes. Human beings, in contrast, can leverage the semantic knowledge (e.g., textual descriptions) to learn the novel concepts that ones have never seen before. Such the "learning to learn" ability inspires the recent study of zero-shot learning (ZSL) [25], which targets at identifying novel classes without any training examples. In practice, the ZSL is achieved via inferring the intermediate semantic representations that may be shared both by the seen and unseen concepts. In particular, the middle-level semantic representations (e.g. attributes) are utilized to make connections between the low-level visual features and high-level class concepts.

Many different semantic representations have been investigated, such as semantic attributes [17], word vectors [20] and gaze embeddings [15]. Though they have to be manually labeled, semantic attributes have been most widely used due to the good merits of "name-ability" and "discriminativeness". Additionally, the attributes can also facilitate the zero-shot generation (ZSG) [41, 30, 43, 19], which aims to generate the images of unseen classes with novel semantic representations.

The image datasets annotated with attributes such as Caltech-UCSD Birds-200-2011 (CUB) [37], SUN Attributes (SUN) [40], aPascal&aYahoo (aP&aY) [7] and Animals with Attributes (AwA) [7], are widely used as the testbed for ZSL algorithms. However, the total number of images and attributes of these dataset are too limited to train from the scratch the state-of-the-art deep models *namely*, VGGs [5], ResNets [13] and DenseNets [14].

Furthermore, there exist several additional issues with these attribute datasets. (1) The categories and images of these datasets may be highly related to ImageNet dataset (used in ILSVRC 2010/2012). In ZSL scenario, it is thus less desirable to directly utilize the deep models pre-trained on ILSVRC 2010/2012 as the feature extractors, which may include the images of novel unseen classes from these datasets [39]. (2) These datasets (CUB and SUN) may focus on each specific visual domain; and yet the datasets for the common object (aP/aY) and animal (AwA) domains do not really have sufficient fine-grained classes to validate the knowledge transfer in zero-shot scenario.

Additionally, there exists serious "co-occurrence bias"

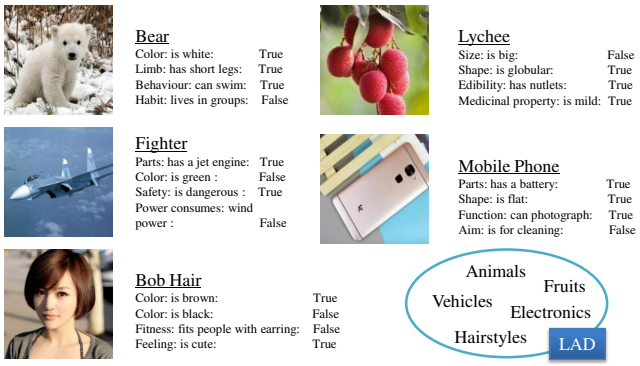


Figure 1. The overview of the proposed LAD dataset. It includes 230 classes belonging to five super-classes (domains). Labels, bounding boxes and attributions are annotated. The upper two attributes are visual attributes, while the bottom two are semantic attributes.

in these datasets. For example, the "person" objects occur in some AwA classes (e.g. "ox" and "horse") with a high frequency ($> 35\%$). Although the co-occurrence of objects may be interesting in some multi-object detection/segmentation tasks, it is not suitable for the single-object learning (classification) task, and it will cause the mis-learning of a particular concept. Such correlation may be implicitly learned and utilized as the cues of identifying zero-shot classes. Specifically, suppose we want to identify two unseen classes – lion and dog. Essentially, these two classes share many common attributes, such as "four legs", "has fur" and so on. Actually, even visually some kind of dog (*e.g.*, Tibetan mastiff) is very similar to lion. However, the zero-shot algorithms may easily identify the dog class by only detecting whether "person" objects are present in the image since high proportion of "person" and "dog" objects are co-occurrence in the dog class of this dataset. Such "co-occurrence" is caused by the way of how we construct the dataset; and thus can be taken as one type of bias. The algorithms implicitly utilize this correlation may be limited to generalize to other domains which do not have such type of correlation.

To alleviate these problems of existing datasets, we are making efforts of contributing the new attribute dataset — Large-scale Attribute Dataset (LAD) to the community. We design a novel label list and collect images from different sources, in order to get more new classes and images different from existing datasets. Except for low-level visual attributes (*e.g.* colors, sizes, shapes), we also provide many attributes about semantic and subject visual properties [9]. For example, as illustrated in Fig. 1, we annotate attributes of diets and habits for "animals"; edibility and medicinal property for "fruits"; safety and usage scenarios for "vehicles"; functions and usage mode for "electronics"; human feelings for "hairstyles". We cluster classes into several super-classes. Each super-class can be viewed as a fine-grained subset, and the attributes are designed for each super-class. Then, the knowledge transfer between fine-grained classes are feasible.

To break the co-occurrence among objects, we collect the images with only single (foreground) object. Overall, we constructed a new attribute dataset which contains 78,017 images from 230 classes. These classes are from 5 different visual domains (super-class), including animals, fruits, vehicles, electronics, and hairstyles. 359 visual, semantic and subjective attributes are annotated for randomly selected 20 images per class.

There are three main contributions of our paper:
1) Present a big dataset: More than 10,000 person-hour time is devoted to the construction of this dataset , which is the larger than the sum of the four most popular datasets in ZSL. Without big data, deep learning models cannot be trained independently on the dataset. The use of pre-trained feature extractors may lead to an impure or unfair competition of different methods.
2) Raise a new problem: The co-occurrence bias problem of existing ZSL datasets is proposed, and its influence on the learning of a concept is well investigated by extensive experiments. The co-occurrence bias may cause the problem of the mis-learning of a concept, which should be prevented during the dataset construction.
3) Provide a better testbed: We provide the re-implementation of seven state-of-the-art methods on our dataset. These results are good baselines for comparison. In addition, Zero-shot Learning Competition of AI Challenger[1], has been organized for promoting ZSL research based on this dataset. More than 110 teams attended this competition. The code and data are provided in Github[2].

2. Related Work

2.1. Zero-shot Learning

In this paper we focus on two zero-shot learning tasks, *namely*, zero-shot recognition and zero-shot generation. Please refer to [11] for a more detailed review.

Zero-shot Recognition (ZSR). ZSR has attracted significant research attention in the past few years. Extensive efforts and previous works can be roughly divided into three groups. (1) Direct embedding from visual space to semantic space (or reverse embedding) [25, 17]. It learns a mapping function from the visual feature space to the semantic embedding space by auxiliary training data; the learned mapping function is directly applied to project the unseen testing images into semantic space and match against the prototype of the novel class/concept. (2) Learning the joint embedding space [1, 38, 31]. Both the image features and

[1]Competition Website: https://challenger.ai/
[2]Github Repository: https://github.com/PatrickZH/A-Large-scale-Attribute-Dataset-for-Zero-shot-Learning

semantic embeddings are jointly projected into a new embedding space. For each given testing unseen image, its label is predicted according to the distance to unseen semantic embeddings in the new space. (3) Transferring structural knowledge from semantic space to visual space [23, 4, 45]. The structural knowledge is learned in semantic space, and then transferred to the visual space for synthesizing the visual instances or classifiers of unseen classes.

Zero-shot Generation (ZSG). In recent years, zero-shot generation methods synthesize images conditioned on attributes/texts using generative models. [41] presented a successful trial to synthesize natural images of birds and faces. They choose Conditional Variational Auto-Encoder as the basic model and then disentangle the foreground and background by introducing a layer representation. PixelCNN is utilized to model images conditioned on labels, tags or latent embeddings [36]. However, new images can be synthesized only based on existing labels, latent embeddings or the linear interpolations of them. Some methods [30, 43, 44, 27] based on conditional GAN [22] have been proposed to generate images with unseen attribute/text representation. In [30], the encoding of text is used as the condition in both the generator and discriminator by concatenating the random noise and image feature maps. [43] proposed a novel model to synthesize and edit facial images. The semi-latent facial attribute space, which includes both learned latent attributes and user-defined attributes, are leveraged as the conditional input of GAN. Note that most of these methods focus on the image generation with particular visual attributes or descriptions, *e.g.* colors, parts of faces, flowers and birds. However, in this paper, we try a more difficult task, *i.e.,* manipulating semantic attributes and generating images with abstract attributes as discussed in Sec. 5.3.

2.2. Image-based Attribute Datasets

Several datasets are repurposed by annotating attributes in order to evaluate the zero-shot learning algorithms. These datasets include Caltech-UCSD Birds-200-2011 (CUB) [37], SUN Attributes (SUN) [40], aPascal&aYahoo (aP&aY) [7], Animals with Attributes (AwA) [7], Public Figures Face Database (PubFig) [16], Human Attributes (HAT) [32] and Unstructured Social Activity Attribute (USAA) [8]. Essentially, any dataset, if labeled with attributes or word vectors, can be used to evaluate the ZSR/ZSG algorithms. The statistics of the most popular four attribute datasets are shown in Tab. 1.

As aforementioned, existing benchmarks have three main drawbacks. (1) The categories and images of existing attribute datasets may be highly reused in ILSVRC 2010/2012 which are frequently used to pre-train the deep feature extractors. Hence, the image feature extractors may have seen many testing ("unseen") classes. (2) The categories are not fine-grained enough. For example, aP/aY contains only 32 coarse-grained categories. As aforementioned, those datasets do not have sufficient fine-grained classes to validate ZSL methods. (3) There exists serious co-occurrence bias in these datasets. AwA and aP/aY contain images with multiple foreground objects; however every image only has a single label. Some objects have a biased co-occurrence with others in particular classes. For instance, 30% classes in AwA have more than 10% images containing "person". Even, in "ox" and "horse" classes, the co-occurrence ratio is greater than 35%. To overcome these three drawbacks, we introduce a new benchmark as the testbed of zero-shot learning.

3. Dataset Construction

The construction process of LAD can be divided into four steps, *namely*, the definition of classes and attributes (Sec. 3.1), image crawling (Sec. 3.2), data preprocessing (Sec. 3.3) and data annotation (Sec. 3.4).

3.1. Definition of Classes and Attributes

Classes. It is of central importance to well define the classes and attributes of an attribute dataset. In general, we expect the LAD have more common classes, and yet fewer shared classes with ImageNet dataset (ILSVRC 2010/2012) [6]. It is nontrivial, since ImageNet dataset is built upon the well-known concept ontology – WordNet [21]. Critically, we define the classes from five domains (super-classes), *namely*, animals, fruits, vehicles, electronics and hairstyles. Animals and fruits are natural products, while vehicles and electronics are artificial products. We choose 50 popular classes for animals, fruits, vehicles, electronics. The hairstyle super-classes include the 30 mostly popular Asian and Western hairstyles. All these classes are selected as less overlapped with the WordNet ID of ILSVRC 2010/2012 dataset. In particular, some classes (*e.g.*, the "fauxhawk" and "mullet" in the hairstyle super-class) have only recently been collected and annotated to the community [42]. Some example images of different classes are shown in the tree structure of Fig. 2.

Attributes. Considering the huge diversity of LAD, we design the attribute list for each super-class; more specifically, we define 123, 58, 81, 75 and 22 attributes for animals, fruits, vehicles, electronics and hairstyles respectively. The defined attributes include visual information (*e.g.* color, shape, size, appearance, part, and texture), the visual semantic information such as "whether an type of animal eats meat?", and subjective visual properties [9], *e.g.*, "whether the hairstyle gives the feeling of cute?". Such a type of attribute definition will facilitate designing zero-shot learning algorithms by transferring various information – visual information, semantic information and subjective visual prop-

	LAD	CUB-bird	SUN	aP/aY	AwA
Images	**78,017**	11,788	14,340	15,339	30,475
Classes	230	200	**717**	32	50
Bounding Box	Yes	Yes	No	Yes	No
Attributes	**359**	312	102	64	85
Annotation Level	20 ins./class	instance	instance	instance	class

Table 1. Statistics and comparison of different datasets used in zero-shot learning.

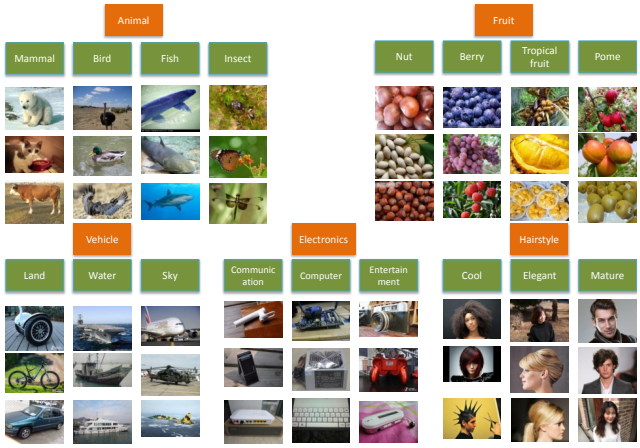


Figure 2. The hierarchical structure of super-classes (domains) and some example classes. Each image represents a class.

erties. Additionally, even more wider types of attributes have been considered here; for example, we annotate attributes of diets and habits for the animal super-class, edibility and medicinal property for the fruit super-class, safety and usage scenarios for the vehicle super-class, functions and usage mode for the electronic super-class. The knowledge of such attributes is referring to *Wikipedia*.

3.2. Image Crawling

We gather the images of each defined class by using the popular search engines, *e.g.* *Baidu* and *Google*. Specifically, to efficiently search enough images, the class names by different languages (*e.g.* English and Chinese) have been used in the search engines. We also use synonyms and determiners to obtain better search results. By this mean, for each class we crawled about 1,000 images with public licenses.

3.3. Preprocessing

Initial Preprocessing. The raw crawled images are very noisy. Huge human efforts are devoted to clean up the crawled images. Specifically, we manually remove those images of low quality (*e.g.*, low-resolution, or large watermark). Also for each class, those duplicated or unrelated noisy images are also manually pruned.

Removing Co-occurrence Bias. Considering that the co-occurrence bias of one dataset mostly comes from the co-existent objects with the proposed object class. For example, many images of the animals in AwA contain the "person" object. To avoid such cases, we prefer the images with iconic view of each class. Particularly, we take as the background, the sky, lakes, land, trees, buildings, blur objects and tiny objects; and those images have more than one foreground object of iconic view would be discarded.

3.4. Annotation

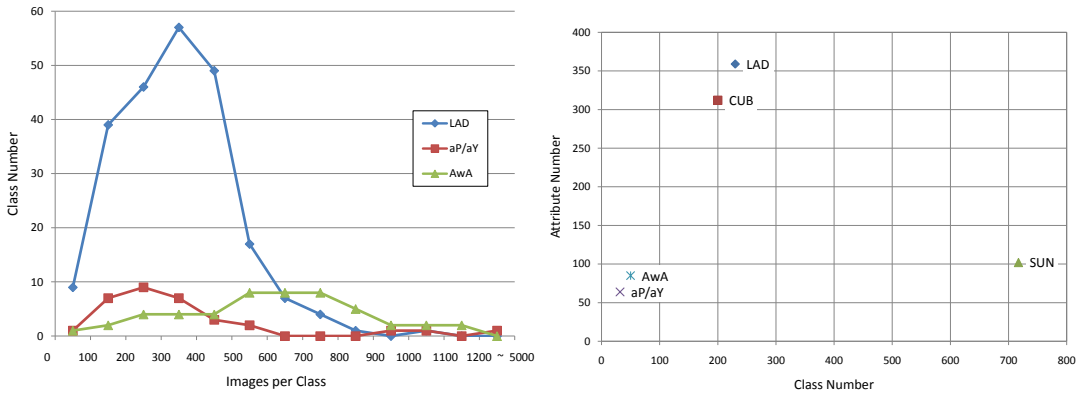
Class Annotation. We also need to further annotate the preprocessed images. In particular, we remove those images whose foreground objects mismatch the class name/label. We finally obtain $78k$ images of all five super-classes as shown in Tab. 1. We also annotate the bounding box of each foreground object.

Attribute Annotation. According to the attribute list defined in Sec. 3.1, we annotate instance-level attributes for selected images. Specifically, we randomly select 20 images per class to annotate attributes. The class-level attributes can be computed as the mean values of the attributes of 20 images.

3.5. Statistics

Total Images. Our LAD dataset contains 78,017 images. As shown in Fig. 3(a), we compare the distribution of image number per class with the AwA and aP/aY datasets. It shows that most classes of LAD have 350 images whilst the AwA and aP/aY have around 650 and 250 images per classes respectively. In particular, AwA has 30,475 images from 50 animals with 85 class-level attributes; and aP/aY includes 15,339 images of 32 classes with 64 instance-level attributes. The area under the curve indicates the total images of each dataset. It means that our LAD is much larger than the AwA and aP/aY datasets.

Classes and Attributes. Fig. 2 provide a hierarchical view of part of classes in our dataset. We can find that every super-class is fine-grained. We also compare the total class and attribute numbers of LAD, CUB, SUN, aP/aY and AwA datasets in Fig. 3(b). Our LAD contains the 359 attributes which is much larger than the attribute number of other datasets. The sheer volume of annotated attributes essentially provide a good testing bed for the zero-shot learning



(a) Statistics of image numbers per class.

(b) Statistics of class and attribute numbers.

Figure 3. Statistics of image, class and attribute numbers.

Dataset	aP/aY	AwA	LAD
Ratio	86.96	76.00	**43.48**

Table 2. The ratios (%) of shared classes between different datasets and ILSVRC 2012. Clearly, our dataset has the lowest overlap ratio.

algorithms. Furthermore, we also introduce the subjective attributes of human hairstyle classes as illustrated in Fig. 1. Comparably, the SUN dataset has 717 classes and yet only 14,340 images and 102 annotated attributes. The CUB-200 2011 bird dataset has 312 attributes, which however only focus on the visual information such as colors, shapes, patterns, sizes and lengths of the birds. In contrast, our LAD introduces the attributes of visual semantic information and subjective visual properties [9]; and we argue that these attributes are much richer semantic representation and potentially can be better used for knowledge transfer in testing the zero-shot algorithms.

We calculate the ratio of shared classes between different attribute datasets and ImageNet dataset. In particular, we use the competition data in ILSVRC 2012, because most deep feature extractors are trained on ILSVRC 2012. For each dataset, we count the number of class names which exist in the WordNet ID of ILSVRC 2012. The ratio is calculated by dividing the total class number of each dataset. As shown in Tab. 2, our dataset has only 43.48% overlap ratio with ILSVRC 2012 which is significantly lower than AwA (76.00%) and aP/aY (86.96%).

4. Data Split

The split of seen/unseen classes significantly influences the performance of zero-shot learning methods. In previous datasets such as CUB, SUN, aP/aY and AwA, only one split of seen/unseen classes is specified for testing zero-shot algorithms. However, due to the distinctive correlations of the classes, it is not reliable nor convincing to evaluate the performance of algorithms on the only one split. Hence, we propose a set of splits of seen/unseen classes for zero-shot learning on our dataset. We adopt the idea of five-fold cross validation to split the seen/unseen classes. Specifically, we shuffle these classes and divide them into 5 folds. Each fold includes 20% classes of every super-class. Each fold is used as the unseen classes (20%) in some split, and the rest folds are seen classes (80%) in the split. In this way, we obtain 5 random splits of seen/unseen classes to evaluate the performance of zero-shot learning on our dataset.

We advocate others to evaluate their ZSL methods on each super-class individually. It means that the data (images, labels, attributes) of each super-class should be used separately. The performance on each super-class should be the average value on all 5 splits. For easy comparison, the average recognition accuracy on all super-classes can be used as the general performance on our dataset. In experiments, we will provide the evaluation of seven state-of-the-art ZSL methods using these splits under the inductive setting.

For supervised learning, we randomly select 70% data from each class as training (train+validation) data and the rest 30% are testing data. These splits will be released along with our dataset.

5. Methods and Experiments

This section will compare the state-of-the-art methods and conduct the experiments under different settings on our dataset. In particular, we consider the supervised learning (Sec. 5.1), zero-shot learning (Sec. 5.2), zero-shot generation (Sec. 5.3). The data and code have been released.

5.1. Supervised Learning

Though our LAD is designed as the testbed for zero-shot learning, we can still validate the LAD in the standard supervised setting. In particular, we provide several baseline results of the supervised learning of objects and attributes.

	w/o. Pre-training		w. Pre-training	
	ResNet	Inception-v3	ResNet	Inception-v3
AwA	49.97	46.26	86.51	87.29
LAD	66.78	44.08	84.92	79.52

Table 3. Object recognition accuracies (%) on two datasets. w. means "with", and w/o. means "without". Clearly, the pre-training on ILSVRC 2012 brings larger performance increase (averagely 38.79%) for AwA than our LAD (averagely 26.79%). This result also denotes that AwA shares more classes with ILSVRC 2012.

In each class, we have labeled training data and unlabeled testing data (the split refers to Sec. 4). We also show that our LAD is large enough and has sufficient images to train the state-of-the-art deep architecture – ResNet.

5.1.1 Object Recognition

We use the state-of-the-art object recognition models, *namely*, Inception-V3 [35] and ResNet [12] to recognize objects. We train the two models under two settings, *namely*, with pre-training on ILSVRC 2012 and without pre-training.

The recognition accuracies of LAD and AwA datasets are shown in Tab. 3. In terms of deep models, ResNet works better than Inception-V3 in most settings. Note that there are 230 classes in LAD and 50 classes in AwA. The chance levels on the two datasets are 0.43% and 2% respectively. However, the recognition accuracy on LAD is close to, even higher than, that on AwA. This phenomenon hints that more images and classes are beneficial to train deep models.

Generally speaking, the pre-training brings significant improvement of recognition accuracies for both two datasets and two models. Averagely, on AwA dataset, the pre-training brings 38.79% increase of recognition accuracy. However, the increase is only 26.79% on our LAD dataset. This gap means that AwA dataset shares more classes with ILSVRC 2012 dataset.

5.1.2 Attribute Recognition

We also consider the task of recognizing different attributes. We use the Inception-v3 features (without fine-tune) to learn attributes. The images that belong to each class are randomly split into 70% training and 30% testing data. The class-level attributes are binarized to be 0 or 1, then the attribute recognition is a binary classification task. We train the Support Vector Machine (SVM) with Multilayer Perception Kernel to learn each attribute of each super-class. The classification accuracy of each attribute is reported as the metric of the performance of attribute recognition.

We histogram the attributes recognition accuracies into several intervals: [0%, 50%), [50%, 60%), [60%, 70%), [70%, 80%), [80%, 90%), [90%, 100%]. As shown in Fig. 4, the recognition accuracies of most attributes of LAD are between 50% and 80%. There are 28 attributes with lower than 50% recognition accuracy (even lower than the chance level), which means those attributes are not well learned. We list some of those attributes in Tab. 4. It is clear that most of those hard attributes are about high-level semantics or those features that can not be visually predicted. For example, habit of animals, safety of vehicles, aim of electronics and feeling of hairstyles are about semantics which is hard to learn. Some attributes, e.g. appearance_has soft skin of animals, hardness_is soft of fruits, material_is made of plastic and sound_is quiet, are also low-level perceptions, but other than vision. Thus the attributes annotated in LAD are multi-modal. Currently, those attributes are difficult to be predicted based on visual perception.

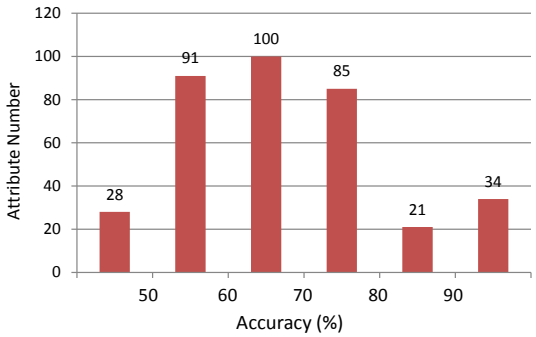


Figure 4. The statistics of attribute recognition accuracies.

5.2. Zero-shot Recognition by Attributes

We propose LAD as the new testbed for zero-shot recognition. In particular, as the sanity check, seven state-of-the-art zero-shot learning algorithms are re-implemented; and their results are reported and compared in this section. We use the data split proposed in Sec. 4. For all methods, we use the ResNet feature extractor which is trained on training images of ILSVRC 2012. We follow the inductive learning setting, *i.e.*, data from unseen classes are not available for training.

Methods. We compare seven state-of-the-art zero-shot learning methods, including SOC [25], ConSE [24], ESZSL [31], SJE [2], SynC [4], LatEm [38], MDP [45]. SOC [25] learns the mapping from image features to semantic output codes (semantic embeddings) using seen classes. Then the learned mapping is used for predicting the semantic output codes of images from unseen classes. ConSE [24] maps the images into the semantic embedding space by the convex combination of the class label embedding vectors. The benefit is that this method does not need an extra training for unseen classes. ESZSL [31] learns the bi-linear mapping function which maps both the image features and semantic embeddings to the new space. SJE [2] proposes to

Super-class	hard-Attributes	Explanation
Animal	diet_eats meat	Whether the animal eats meat
	habit_is nocturnal	whether the animal's habit is nocturnal
Fruits	growth_grow on trees	whether the fruit grows on the tree
	hardness_is soft	whether the hardness of the fruit is soft
Vehicles	safety_is safe	whether the vehicle is safety
	material_is made of plastic	whether the vehicle is made of plastic
Electronics	aim_is for display	whether the electronics is designed for display
	sound_is quiet	whether the sound of electronics is quite
Hairstyles	feeling_Is elegant	whether the hairstyle gives the feeling of elegant
	feeling_Is sexy	whether the hairstyle gives the feeling of sexy

Table 4. Some example attributes with low recognition accuracies.

	SOC	ConSE	ESZSL	SJE	SynC	LatEm	MDP
Animals	50.76	36.87	50.15	61.89	61.60	**63.92**	62.16
Fruits	40.01	29.77	37.23	46.39	51.42	44.23	**56.40**
Vehicles	56.98	37.48	45.75	63.00	54.89	60.94	**65.09**
Electronics	33.73	28.27	32.83	39.51	42.97	40.71	**45.11**
Hairstyles	**42.45**	24.55	31.84	38.50	29.10	38.53	42.12
Average	44.79	31.39	39.56	49.86	48.00	49.67	**54.18**

Table 5. The performance (%) of seven state-of-the-art ZSR methods on our dataset.

learn the compatibility function which measures the compatibility between the the image features and semantic embeddings. The function is trained on seen classes and tested on unseen classes. SynC [4] learns to synthesize classifiers for unseen classes by the linear combination of classifiers for seen classes. LatEm [38] proposes a new compatibility function which is a collection of bilinear maps. These bilinear maps can discover latent variables. MDP [45] aims to learn the local structure in the semantic embedding space, then transfer it to the image feature space. In the image feature space, the distribution of unseen classes are estimated based on the transferred structural knowledge and the distribution of seen classes.

We can roughly classify these methods into three groups in term of how the knowledge is transferred (refer to Sec. 2). The first group includes SOC [25], ConSE [24] and ESZSL [31], while SJE [2] and LatEm [38]) belong to the second group. The third group contains SynC [4] and MDP [45].

Note that there are many zero-shot algorithms such as SS-Voc [10] that heavily rely on the word vectors (e.g. Word2Vec[20], or GloVec[26]) of the class names. However, in LAD, the class name is less informative to represent the whole data distribution in the semantic layer, *e.g.*, the "fauxhawk" class in the hairstyle super-class. Therefore, for a more fair comparison, the algorithms that are heavily relying on the word vectors have not been compared here. The zero-shot recognition is conducted on each super-class separately.

Tab. 5 shows the zero-shot recognition accuracies of different methods. In general, MDP achieves the best performance (54.18% averagely). This result is higher than the runner-up (SJE) by 4.32%. Among all the super-classes, most algorithms can achieve relative high performance on the super-classes of "Animals" and "Vehicles". This is reasonable, since these two super-classes include very common objects/concepts which have been widely collected in the ImageNet dataset. Hence, the feature extractor pre-trained on ILSVRC 2012 may work better on the two super-classes than the other ones. Almost all the algorithms have relative low performance on "Hairstyles" super-class, even although we only split 6 unseen classes. More impressively, the SOC [25] proposed in 2009 can beat all the other methods on "Hairstyles" super-class.

5.3. Zero-shot Generation

We also conduct experiments for the zero-shot generation task. This task aims to generate images of unseen classes, *i.e.*, those with novel attribute representations. Note that this task is extremely challenging due to both the diverse and fine-grained classes in LAD and the high-semantic multi-modal. In particular, we conduct the experiments on the "Animals" super-class.

Methods. Based on the Deep Convolutional Generative Adversarial Networks (DCGAN) [28], we introduce the condition by concatenating the condition vector and the noise vector. In DCGAN, the generator and discriminator are two deep convolutional networks which have 4 convolution layers. Refer to [28] for the detailed convolutional structure. Both input and output images are reshaped to $64 \times 64 \times 3$. The attributes of each image serves as the condition. In this way, the learned DCGAN can generate images conditioned on attributes. We illustrate the model

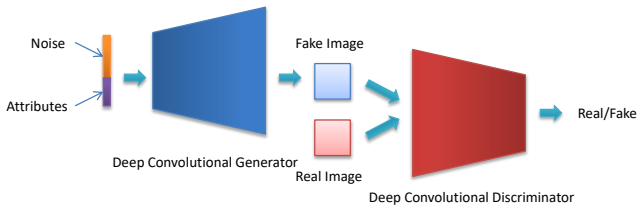


Figure 5. The model structure for zero-shot generation.

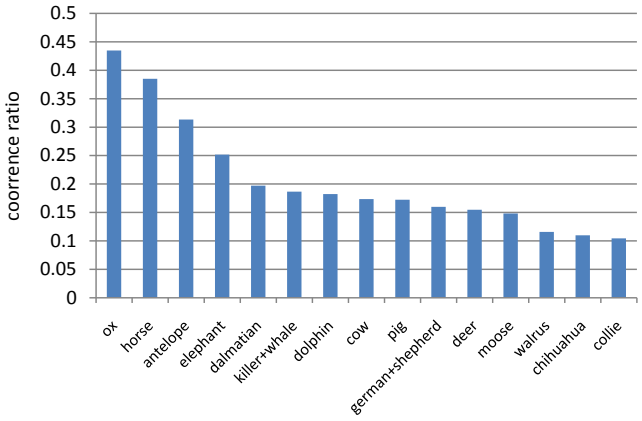


Figure 6. The co-occurrence bias on AwA dataset. This figure shows the top 15 animal classes with high co-occurrence ratios of persons.

structure in Fig. 5.

Results. The animals in our dataset are divided into 3 folds in term of their attributes – "can swim"(1/0) and "can fly" (1/0), namely, 00, 01, 10. Then we train a GAN conditioned on the two attributes. The trained model is used to generate new images with both seen attribute representations and the unseen one (with the attribute representation "11"). As shown in Fig. 7, objects with the particular seen attributes can be generated. From the left to right, more clear images are generated with more training iteration. The objects in the first three rows look like "monkey", "fish" and "bird" respectively. In the 4th row, the generated unseen object looks like a "fish" in the 2nd stage (column). Later, the "wings" are observed in the 3rd stage. This result means that the novel object can be generated based on novel the attribute representation. Note that the distribution of the training seen images in LAD is very diverse; and thus it is intrinsically a very challenging task to generate the images of novel unseen classes.

Analysis of Co-occurrence Bias in AwA. We also study the co-occurrence bias in previous datasets. We analyze and visualize the influence of the co-occurrence bias on the learning of a particular concept. Specifically, we first count the co-occurrence of person and different animals in AwA dataset. We use YOLO [29], which is pre-trained on MS-COCO [18], to detect person in each image. We only count

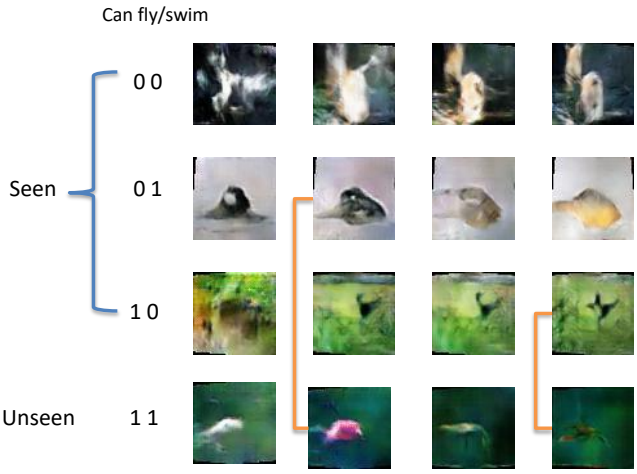


Figure 7. Results of zero-shot generation. The seen classes have the attributes 00, 01 and 10, which denote whether it can fly or swim. The unseen class has the attribute representation of 11. From left to right, we display the generated images with more training iterations.

Figure 8. Visualization of the co-occurrence bias in AwA. The upper three images are synthesized images of "ox", and the below images are those of "horse". Blue boxes are animals, while red boxes are "persons", i.e., bias.

those appeared person with high ($> 70\%$) probability. As shown in Fig. 6, 15 classes in AwA have large ($> 10\%$) co-occurrence ratio of person. Clearly, the co-occurrence bias of person is serious in AwA. To visualize the influence of the co-occurrence bias, we use the GAN, which can well capture the data distribution, to learn the concepts "ox" and "horse". Fig. 8 illustrates the synthesized images of the two animals. Except animals in blue boxes, the "persons" in red boxes are also synthesized in the image. These synthesis results illustrate that the co-occurrence bias may cause the mis-learning of a particular concept. Thus during the construction of LAD, we reduce the correlation bias by filtering multi-object images, *i.e.*, we preserve images with only one foreground object.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 819–826, 2013.

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015.

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

[4] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, pages 5327–5336, 2016.

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[7] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785, 2009.

[8] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *European Conference on Computer Vision*, pages 530–543. Springer, 2012.

[9] Y. Fu, T. M. Hospedales, T. Xiang, J. Xiong, S. Gong, Y. Wang, and Y. Yao. Robust subjective visual property prediction from crowdsourced pairwise labels. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):563–577, 2016.

[10] Y. Fu and L. Sigal. Semi-supervised vocabulary-informed learning. In *CVPR*, pages 5337–5346, 2016.

[11] Y. Fu, T. Xiang, Y.-G. Jiang, L. Sigal, and S. Gong. Recent advances in zero-shot learning. *IEEE SPM*.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. 2016.

[14] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. In *CVPR*, 2017.

[15] N. Karessli, Z. Akata, B. Schiele, and A. Bulling. Gaze embeddings for zero-shot image classification. In *30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[16] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009.

[17] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 36(3):453–465, 2014.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[19] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, pages 5040–5048, 2016.

[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.

[21] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[22] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[23] S. Naha and Y. Wang. Zero-shot object recognition using semantic label vectors. In *CRV*, 2015.

[24] 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. *NIPS*, 2013.

[25] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, pages 1410–1418, 2009.

[26] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[27] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

[28] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[29] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[30] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

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

[32] G. Sharma and F. Jurie. Learning discriminative spatial representation for image classification. In *BMVC 2011-British Machine Vision Conference*, pages 1–11. BMVA Press, 2011.

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

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Zbigniew-Wojna. Rethinking the inception architecture for computer vision. In *arxiv*, 2015.

[36] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016.

[37] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology, 2011.

[38] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, pages 69–77, 2016.

[39] Y. Xian, B. Schiele, and Z. Akata. Zero-shot learning - the good, the bad and the ugly. In *CVPR*, 2017.

[40] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.

[41] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016.

[42] W. Yin, Y. Fu, Y. Ma, Y.-G. Jiang, T. Xiang, and X. Xue. Learning to generate and edit hairstyles. In *ACM MM*, 2017.

[43] W. Yin, Y. Fu, L. Sigal, and X. Xue. Semi-latent gan: Learning to generate and modify facial images from attributes. *arXiv preprint arXiv:1704.02166*, 2017.

[44] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1612.03242*, 2016.

[45] B. Zhao, B. Wu, T. Wu, and Y. Wang. Zero-shot learning posed as a missing data problem. In *Proceedings of ICCV Workshop*, pages 2616–2622, 2017.