Yanwei Fu, Tao Xiang, Yu-Gang Jiang,
Xiangyang Xue, Leonid Sigal, and Shaogang Gong

Recent Advances in Zero-Shot Recognition

Toward data-efficient understanding of visual content

With the recent renaissance of deep convolutional neural networks (CNNs), encouraging breakthroughs have been achieved on the supervised recognition tasks, where each class has sufficient and fully annotated training data. However, to scale the recognition to a large number of classes with few or no training samples for each class remains an unsolved problem. One approach is to develop models capable of recognizing unseen categories without any training instances, or zero-shot recognition/learning. This article provides a comprehensive review of existing zero-shot recognition techniques covering various aspects ranging from representations of models, data sets, and evaluation settings. We also overview related recognition tasks including one-shot and open-set recognition, which can be used as natural extensions of zero-shot recognition when a limited number of class samples become available or when zero-shot recognition is implemented in a real-world setting. We highlight the limitations of existing approaches and point out future research directions in this existing new research area.

Introduction

Humans can distinguish at least 30,000 basic object categories and many more subordinate ones (e.g., breeds of dogs). They can also create new categories dynamically from a few examples or purely based on high-level descriptions. In contrast, most existing computer vision techniques require hundreds, if not thousands, of labeled samples for each object class to learn a recognition model. Inspired by humans' ability to recognize objects without first seeing examples, the research area of learning to learn, or lifelong learning [1], has received increasing interest.

These studies aim to intelligently apply previously learned knowledge to help future recognition tasks. In particular, a major topic in this research area is building recognition models capable of recognizing novel visual categories that have no associated labeled training samples (i.e., zero-shot learning), few training examples (i.e., one-shot learning), and recognizing the visual categories under an "open-set" setting where the testing instance could belong to either seen or unseen/novel categories.

Digital Object Identifier 10.1109/MSP.2017.2763441
Date of publication: 9 January 2018

These problems can be solved under the setting of transfer learning. Typically, transfer learning emphasizes the transfer of knowledge across domains, tasks, and distributions that are similar but not the same. Transfer learning refers to the problem of applying the knowledge learned in one or more auxiliary tasks/domains/sources to develop an effective model for a target task/domain.

To recognize zero-shot categories in the target domain, one has to utilize the information learned from source domain. Unfortunately, it may be difficult for existing methods of domain adaptation to be directly applied to these tasks, since there are only few training instances available on the target domain. Thus, the key challenge is to learn domain-invariant and generalizable feature representation and/or recognition models usable in the target domain.

Overview of zero-shot recognition

Zero-shot recognition can be used in a variety of research areas, such as neural decoding from functional magnetic resonance imaging [2], face verification [3], object recognition [4], and video understanding [5]–[7]. The tasks of identifying classes without any observed data is called *zero-shot learning*. Specifically, in the settings of zero-shot recognition, the recognition model should leverage training data from source/auxiliary data set/domain to identify the unseen target/testing data set/domain. Thus, the main challenge of zero-shot recognition is how to generalize the recognition models to identify the novel object categories without accessing any labeled instances of these categories.

The key idea underpinning zero-shot recognition is to explore and exploit the knowledge of how an unseen class (in the target domain) is semantically related to the seen classes (in the source domain). We explore the relationship of seen and unseen classes in the section "Semantic Representations in Zero-Shot Recognition," through the use of intermediate-level semantic representations. These semantic representations are typically encoded in a high-dimensional vector space. The common semantic representations include semantic attributes (see the section "Semantic Attributes") and semantic word vectors (see the section "Semantic Representations Beyond Attributes"), encoding linguistic context. The semantic representation is assumed to be shared between the auxiliary/source and target/test data set. Given a predefined semantic representation, each class name can be represented by an attribute vector or a semantic word vector—a representation termed *class prototype*.

Because the semantic representations are universal and shared, they can be exploited for knowledge transfer between the source and target data sets (see the section "Models for Zero-Shot Recognition"), to enable the recognition of novel, unseen classes. A projection function mapping visual features to the semantic representations is typically learned from the auxiliary data, using an embedding model (see the section "Embedding Models"). Each unlabeled target class is represented in the same embedding space using a class "prototype." Each projected target instance is then classified, using the recognition model, by measuring the similarity of projection to the class prototypes in the embedding space (see the section "Models for Zero-Shot Recognition"). Additionally, under an open-set setting, where the test instances could belong to either the source or target categories, the instances of target sets can also be taken as outliers of the source data; therefore, novelty detection [8] needs to be employed first to determine whether a testing instance is on the manifold of source categories and, if it is not, it will be further classified into one of the target categories.

Zero-shot recognition can be considered a type of lifelong learning. For example, when reading a description "flightless birds living almost exclusively in Antarctica," most of us know and can recognize that the description refers to a penguin, even though many people probably have not seen a real penguin in person. In cognitive science [9], studies explain that humans are able to learn new concepts by extracting intermediate semantic representation or high-level descriptions (i.e., flightless, bird, living in Antarctica) and transferring knowledge from known sources (other bird classes, e.g., swan, canary, cockatoo, and so on) to the unknown target (penguin). That is the reason why humans are able to understand new concepts with no (zero-shot recognition) or only few training samples (few-shot recognition). This ability is termed *learning to learn*.

Humans can recognize newly created categories from a few examples or merely based on a high-level description, e.g., they are able to easily recognize the video event "Germany World Cup Winner Celebrations 2014," which, by definition, did not exist before July 2014. To teach machines to recognize the numerous visual concepts dynamically created by combining a multitude of existing concepts, one would require an exponential set of training instances for a supervised learning approach. As such, the supervised approach would struggle with the one-off and novel concepts such as "Germany World Cup Winner Celebrations 2014," because no positive video samples would be available before July 2014 when Germany ultimately beat Argentina to win the Cup. Therefore, zero-shot recognition is crucial for recognizing dynamically created novel concepts that are composed of new combinations of existing concepts. With zero-shot learning, it is possible to construct a classifier for "Germany World Cup Winner Celebrations 2014" by transferring knowledge from related visual concepts with ample training samples, e.g., "FC Bayern Munich—Champions of Europe 2013" and "Spain World Cup Winner Celebrations 2010."

Semantic representations in zero-shot recognition

Semantic representations can be categorized into two categories: semantic attributes and beyond. We briefly review relevant papers in Table 1.

Semantic attributes

An attribute (e.g., "has wings") refers to the intrinsic characteristic that is possessed by an instance or a class (e.g., bird) (Fu et al. [5]), or indicates properties (e.g., spotted) or annotations (e.g., has a head) of an image or an object (Lampert et al. [4]). Attributes describe a class or an instance, in contrast to the typical classification, which names an instance. Farhadi et al. [10] learned a richer set of attributes, including parts, shape,

materials, etc. Another commonly used methodology (e.g., in human action recognition (Liu et al. [6]), and in attribute and object-based modeling (Wang et al. [11]) is to take the attribute labels as latent variables on the training data set, e.g., in the form of a structured latent support vector machine (SVM) model where the objective is to minimize prediction loss. The attribute description of an instance or a category is useful as a semantically meaningful intermediate representation bridging a gap between low-level features and high-level class concepts (Palatucci et al. [2]).

The attribute-learning approaches have emerged as a promising paradigm for bridging the semantic gap and addressing data sparsity through transferring attribute knowledge in image and video understanding tasks. A key advantage of attribute learning is that it provides an intuitive mechanism for multitask learning (Hwang et al. [12]) and transfer learning (Hwang et al. [12]). Particularly, attribute learning enables learning with few or zero instances of each class via attribute sharing, i.e., zero-shot and one-shot learning. The challenge of zero-shot recognition is to recognize unseen visual object categories without any training exemplars of the unseen class. This requires the knowledge transfer of semantic information from auxiliary (seen) classes with example images, to unseen target classes.

Later works (Parikh et al. [13]) extended the unary/binary attributes to compound attributes, which makes them extremely useful for information retrieval (e.g., by allowing complex queries such as "Asian women with short hair, big eyes, and high cheekbones") and identification (e.g., finding an actor whose name you forgot, or an image that you have misplaced in a large collection).

In a broader sense, the attribute can be taken as one special type of subjective visual property [14], which indicates the task of estimating continuous values representing visual properties observed in an image/video. These properties are also examples of attributes, including image/video interestingness [15], and human-face age estimation [16]. Image interestingness was studied in Gygli et al. [15], which showed that three cues contribute the most to interestingness: aesthetics, unusualness/novelty, and general preferences; the last of which refers to the fact that people, in general, find certain types of scenes more interesting than others, e.g., outdoor-natural versus indoor-manmade. Jiang et al. [17] evaluated different features for video interestingness prediction from crowdsourced pairwise comparisons. The ACM International Conference on Multimedia Retrieval 2017 published a special issue ("Multimodal Understanding of Subjective Properties") on the applications of multimedia analysis for subjective property understanding, detection and retrieval (see http://www.icmr2017.ro/call-for-special-sessions-s1.php). These subjective visual properties can be used as an intermediate representation for zero-shot recognition as well as other visual recognition tasks, e.g., people can be recognized by the description of how pale their skin complexion is and/or how chubby their face looks [13]. Next, we will briefly review different types of attributes.

Table 1. Different types of semantic representations for zero-shot recognition.

Different Types of Attributes	Papers
User-defined attributes	[4], [5], [11], [18]–[21], [23], [24], [30]
Relative attributes	[13], [14], [25]–[29]
Data-driven attributes	[5]–[7], [10], [31]–[33]
Video attributes	[34]–[38]
Concept ontology	[39]–[42]
Semantic word embedding	[8], [43]–[48]

User-defined attributes

User-defined attributes are defined by human experts [4] or by concept ontology [5]. Different tasks may also necessitate and contain distinctive attributes, such as facial and clothes attributes [11], [18]–[20], attributes of biological traits (e.g., age and gender) [21], product attributes (e.g., size, color, price), and three-dimensional shape attributes [22]. Such attributes transcend the specific learning tasks and are, typically, prelearned independently across different categories, thus allowing transference of knowledge [23]. Essentially, these attributes can either serve as the intermediate representations for knowledge transfer in zero-shot, one-shot, and multitask learning, or be directly employed for advanced applications, such as clothes recommendations [11].

Ferrari et al. [24] studied some elementary properties such as color and/or geometric pattern. From human annotations, they proposed a generative model for learning simple color and texture attributes. The attribute can be viewed as either unary (e.g., red color, round texture), or binary (e.g., black/white stripes). The unary attributes are simple attributes, whose characteristic properties are captured by individual image segments (appearance for red, shape for round). In contrast, the binary attributes are more complex attributes, whose basic element is a pair of segments (e.g., black/white stripes).

Relative attributes

The aforementioned attributes use a single value to represent the strength of an attribute being possessed by one instance/class; they can indicate properties (e.g., spotted) or annotations of images or objects. In contrast, relative information, in the form of relative attributes, can be used as a more informative way to express richer semantic meaning and thus better represent visual information. The relative attributes can be directly used for zero-shot recognition [13].

Relative attributes (Parikh et al. [13]) were first proposed to learn a ranking function capable of predicting the relative semantic strength of a given attribute. The annotators give pairwise comparisons on images, and a ranking function is then learned to estimate relative attribute values for unseen images as ranking scores. These relative attributes are learned as a form of richer representation, corresponding to the strength of visual properties, and used in a number of tasks including visual recognition with sparse data, interactive image search

(Kovashka et al. [25]), semisupervised (Shrivastava et al. [26]), and active learning (Biswas et al. [27], [28]) of visual categories. Kovashka et al. [25] proposed a novel model of feedback for image search where users can interactively adjust the properties of exemplar images by using relative attributes to best match his or her ideal queries.

Fu et al. [14] extended the relative attributes to "subjective visual properties" and proposed a learning-to-rank model of pruning the annotation outliers/errors in crowdsourced pairwise comparisons. Given only weakly supervised pairwise image comparisons, Singh et al. [29] developed an end-to-end deep convolutional network to simultaneously localize and rank relative visual attributes. The localization branch in [29] is adapted from the spatial transformer network.

Data-driven attributes

The attributes are usually defined by extra knowledge of either expert users or concept ontology. To better augment such user-defined attributes, Parikh et al. [31] proposed a novel approach to actively augment the vocabulary of attributes to both help resolve intraclass confusions of new attributes and coordinate the "name-ability" and "discriminativeness" of candidate attributes. However, such user-defined attributes are far from enough to model the complex visual data. The definition process can still be either inefficient (requiring substantial effort from user experts) and/or insufficient (descriptive properties may not be discriminative). To tackle such problems, it is necessary to automatically discover more discriminative intermediate representations from visual data, i.e. data-driven attributes. The data-driven attributes can be used in zero-shot recognition tasks [5], [6].

Despite previous efforts, an exhaustive space of attributes is unlikely to be available due to the expense of ontology creation and the simple fact that semantically obvious attributes for humans do not necessarily correspond to the space of detectable and discriminative attributes. One method of collecting labels for large-scale problems is to use Amazon Mechanical Turk (AMT). However, even with excellent quality assurance, the results collected still exhibit strong label noise. Thus, label-noise [32] is a serious issue in learning from either AMT or existing social metadata. More subtly, even with an exhaustive ontology, only a subset of concepts from the ontology are likely to have sufficient annotated training examples, so the portion of the ontology that is effectively usable for learning may be much smaller. This inspired the works of automatically mining the attributes from data.

Data-driven attributes have only been explored in a few previous works. Liu et al. [6] employed an information-theoretic approach to infer the data-driven attributes from training examples by building a framework based on a latent SVM formulation. They directly extended the attribute concepts in images to comparable "action attributes" to better recognize human actions. Attributes are used to represent human actions from videos and enable the construction of more descriptive models for human action recognition. They augmented user-defined attributes with data-driven attributes to better differentiate existing classes. Farhadi et al. [10] also learned user-defined and data-driven attributes.

The data-driven attribute works in [6] and [10] are limited. First, they learn the user-defined and data-driven attributes separately rather than jointly in the same framework. Therefore, data-driven attributes may rediscover the patterns that exist in the user-defined attributes. Second, the data-driven attributes are mined from data, and we do not know the corresponding semantic attribute names for the discovered attributes. For those reasons, usually data-driven attributes cannot be directly used in zero-shot learning. These limitations inspired the works of [5] and [7]. Fu et al. [5], [7] addressed the tasks of understanding multimedia data with sparse and incomplete labels. Particularly, they studied the videos of social group activities by proposing a novel scalable probabilistic topic model for learning a semilatent attribute space. The learned multimodal semilatent attributes can enable multitask learning, one-shot learning, and zero-shot learning. Habibian et al. [33] proposed a new type of video representation by learning the "VideoStory" embedding from videos and corresponding descriptions. This representation can also be interpreted as data-driven attributes. The work won the Best Paper Award at ACM Multimedia 2014.

Video attributes

Most existing studies on attributes focus on object classification from static images. Another line of work instead investigates attributes defined in videos, i.e., video attributes, which are very important for corresponding video-related tasks such as action recognition and activity understanding. Video attributes can correspond to a wide range of visual concepts such as objects (e.g., animal), indoor/outdoor scenes (e.g., meeting, snow), actions (e.g. blowing out a candle), and events (e.g., wedding ceremony), and so on. Compared to static image attributes, many video attributes can only be computed from image sequences and are more complex in that they often involve multiple objects.

Video attributes are closely related to video concept detection in the multimedia community. The video concepts in a video ontology can be taken as video attributes in zero-shot recognition. Depending on the ontology and models used, many approaches on video concept detection (Chang et al. [49], Gan et al. [42], and Qin et al. [50]) can therefore be seen as addressing a subtask of video attribute learning to solve zero-shot video event detection. Some works aim to automatically expand or enrich the set of video tags [35] given a search query. In this case, the expanded/enriched tagging space has to be constrained by a fixed concept ontology, which may be very large and complex [35]. For example, there is a vocabulary space of more than 20,000 tags in [35].

Zero-shot video event detection has also recently attracted much research attention. The video event is a higher-level semantic entity and is typically composed of multiple concepts/video attributes. For example, a "birthday party" event consists of multiple concepts, e.g., "blowing out a candle" and "birthday cake." The semantic correlation of video concepts has also

been utilized to help predict the video event of interest, such as weakly supervised concepts [51], pairwise relationships of concepts (Gan et al. [42]), and general video understanding by object and scene semantics attributes [36], [37]. Note, a full survey of recent works on zero-shot video event detection is beyond the scope of this article.

Semantic representations beyond attributes

Besides the attributes, there are many other types of semantic representations, e.g., semantic word vector and concept ontology. Representations that are directly learned from textual descriptions of categories, such as Wikipedia articles [52], [53], sentence descriptions [54], or knowledge graphs [40], have also been investigated.

Concept ontology

Concept ontology is directly used as the semantic representation alternative to attributes. For example, WordNet is one of the most widely studied concept ontologies. It is a large-scale semantic ontology built from a large lexical data set of English. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets) that indicate distinct concepts. The idea of semantic distance, defined by the WordNet ontology, is also used by Rohrbach et al. [40] for transferring semantic information in zero-shot learning problems. They thoroughly evaluated many alternatives of semantic links between auxiliary and target classes by exploring linguistic bases such as WordNet, Wikipedia, Yahoo Web, Yahoo Image, and Flickr Image. Additionally, WordNet has been used for many vision problems. Fergus et al. [39] leveraged the WordNet ontology hierarchy to define semantic distance between any two categories for sharing labels in classification. The COSTA [41] model exploits the co-occurrences of visual concepts in images for knowledge transfer in zero-shot recognition.

Semantic word vectors

Recently, word vector approaches based on distributed language representations have gained popularity in zero-shot recognition [8], [43]–[46]. A user-defined semantic attribute space is predefined, and each dimension of the space has a specific semantic meaning according to either human experts or concept ontology (e.g., one dimension could correspond to "has fur," and another "has four legs") (see the section "User-Defined Attributes"). In contrast, the semantic word vector space is trained from linguistic knowledge bases such as Wikipedia and UMBCWebBase using natural language processing models [47]. As a result, although the relative positions of different visual concepts will have semantic meaning, e.g., a cat would be closer to a dog than a sofa, each dimension of the space does not have a specific semantic meaning. The language model is used to project each class' textual name into this space. These projections can be used as prototypes for zero-shot learning. Socher et al. [8] learned a neural network model to embed each image into a 50-dimensional word vector semantic space, which was obtained using an unsupervised linguistic model [47] trained on Wikipedia text. The images from either known or unknown classes could be mapped into such word vectors and classified by finding the closest prototypical linguistic word in the semantic space.

Distributed semantic word vectors have been widely used for zero-shot recognition. The skip-gram model and continuous bag-of-words (CBOW) model [55] were trained from a large scale of text corpora to construct semantic word space. Different from the unsupervised linguistic model [47], distributed word vector representations facilitate modeling of syntactic and semantic regularities in language and enable vector-oriented reasoning and vector arithmetics. For example, Vec("Moscow") should be much closer to Vec("Russia") + Vec("capital") than Vec("Russia") or Vec("capital") in the semantic space. One possible explanation and intuition underlying these syntactic and semantic regularities is the distributional hypothesis [56], which states that a word's meaning is captured by other words that co-occur with it. Frome et al. [46] further scaled such ideas to recognize large-scale data sets. They proposed a deep visual-semantic embedding model to map images into a rich semantic embedding space for large-scale zero-shot recognition. Fu et al. [45] showed that such a reasoning could be used to synthesize all different label combination prototypes in the semantic space and thus is crucial for multilabel zero-shot learning. More recent work of using semantic word embedding includes [43] and [44].

More interestingly, the vector arithmetics of semantic emotion word vectors matches the psychological theories of emotion, such as Ekman's six pan-cultural basic emotions or Plutchik's emotion. For example, Vec("Surprise") + Vec("Sadness") is very close to Vec("Disappointment"); and Vec("Joy") + Vec("Trust") is very close to Vec("Love"). Since there are usually thousands of words that can describe emotions, zero-shot emotion recognition has been also investigated in [48].

Models for zero-shot recognition

With the help of semantic representations, zero-shot recognition can usually be solved by first learning an embedding model (see the section "Embedding Models") and then solving recognition (see the section "Recognition Models in the Embedding Space"). To the best of our knowledge, a general embedding formulation of zero-shot recognition was first introduced by Larochelle et al. [57]. They embedded a handwritten character with a typed representation that further helped to recognize unseen classes.

The embedding models aim to establish connections between seen classes and unseen classes by projecting the low-level features of images/videos close to their corresponding semantic vectors (prototypes). Once the embedding is learned from known classes, novel classes can be recognized based on the similarity of their prototype representations and predicted representations of the instances in the embedding space. The recognition model matches the projection of the image features against the unseen class prototypes (in the embedding space).

Embedding models

Bayesian models

The embedding models can be learned using a Bayesian formulation, which enables easy integration of prior knowledge of each

type of attribute to compensate for limited supervision of novel classes in image and video understanding. A generative model is first proposed by Ferrari and Zisserman in [24] for learning simple color and texture attributes.

Lampert et al. [4] is the first to study the problem of object recognition of categories for which no training examples are available. Direct attribute prediction (DAP) and indirect attribute prediction (IAP) are the first two models for zero-shot recognition [4]. DAP and IAP algorithms use a single model that first learns embedding using an SVM and then does recognition using Bayesian formulation. The DAP and IAP further inspired later works that employ generative models to learn the embedding, including those with topic models [5], [7], [58] and random forests [59]. We briefly describe the DAP and IAP models as follows:

- *DAP model*: Assume the relation between known classes, $y_i, \ldots, y_k$, unseen classes, $z_1, \ldots, z_L$, and descriptive attributes $a_1, \ldots, a_M$ is given by the matrix of binary associations values a_m^y and a_m^z. Such a matrix encodes the presence/absence of each attribute in a given class. Extra knowledge is applied to define such an association matrix—for instance, by leveraging human experts (Lampert et al. [4]), by consulting a concept ontology (Fu et al. [7]), or by semantic relatedness measured between class and attribute concepts (Rohrbach et al. [40]). In the training stage, the attribute classifiers are trained from the attribute annotations of known classes $y_i, \ldots, y_k$. At the test stage, the posterior probability $p(a_m \mid x)$ can be inferred for an individual attribute a_m in an image x. To predict the class label of object class z,

$$p(z \mid x) = \sum_{a \in \{0,1\}^M} p(z \mid a) p(a \mid x) \tag{1}$$

$$= \frac{p(z)}{p(a^z)} \prod_{m=1}^{M} p(a_m \mid x)^{a_m^z}. \tag{2}$$

- *IAP model*: The DAP model directly learns attribute classifiers from the known classes, while the IAP model builds attribute classifiers by combining the probabilities of all associated known classes. It is also introduced as a direct similarity-based model in Rohrbach et al. [40]. In the training step, we can learn the probabilistic multiclass classifier to estimate $p(y_k \mid x)$ for all training classes $y_i, \ldots, y_k$. Once $p(a \mid x)$ is estimated, we use it in the same way as we do for DAP in zero-shot learning classification problems. In the testing step, we predict

$$p(a_m \mid x) = \sum_{k=1}^{K} p(a_m \mid y_k) p(y_k \mid x). \tag{3}$$

Semantic embedding

Semantic embedding learns the mapping from visual feature space to the semantic space which has various semantic representations. As discussed in the section "Semantic Attributes," the attributes are introduced to describe objects, and the learned attributes may not be optimal for recognition tasks. To this end, Akata et al. [60] proposed the idea of label embedding that takes attribute-based image classification as a label-embedding problem by minimizing the compatibility function between an image and a label embedding. In their work, a modified ranking objective function was derived from the WSABIE model [61]. As object-level attributes may suffer from the problems of the partial occlusions and scale changes of images, Li et al. [62] proposed learning and extracting attributes on segments containing the entire object and then joint learning for simultaneous object classification and segment proposal ranking by attributes. They thus learned the embedding by the max-margin empirical risk over both the class label and the segmentation quality. Other semantic embedding algorithms such as semisupervised max-margin learning framework [63], latent SVM [64], or multitask learning [12], [65], [66] have also been investigated.

Embedding into common spaces

Besides the semantic embedding, the relationship of visual and semantic space can be learned by jointly exploring and exploiting a common intermediate space. Extensive efforts [53], [66]–[70] had been made toward this direction. Akata et al. [67] learned a joint embedding semantic space between attributes, text, and hierarchical relationships. Ba et al. [53] employed text features to predict the output weights of both the convolutional and the fully connected layers in a deep CNN.

On one data set, there may exist many different types of semantic representations. Each type of representation may contain complementary information. Fusing them can potentially improve the recognition performance. Thus, several recent works studied different methods of multiview embedding. Fu et al. [71] employed the semantic class label graph to fuse the scores of different semantic representations. Similarly, label relation graphs have also been studied in [72] and significantly improved large-scale object classification in supervised and zero-shot recognition scenarios.

A number of successful approaches to learning a semantic embedding space reply on canonical component analysis (CCA). Hardoon et al. [73] proposed a general kernel CCA method for learning semantic embedding of web images and their associated text. Such embedding enables a direct comparison between text and images. Many more works [74] focused on modeling the images/videos and associated text (e.g., tags on Flickr/YouTube). Multiview CCA is often exploited to provide unsupervised fusion of different modalities. Gong et al. [74] also investigated the problem of modeling Internet images and associated text or tags and proposed a three-view CCA embedding framework for retrieval tasks. Additional view allows their framework to outperform a number of two-view baselines on retrieval tasks. Qi et al. [75] proposed an embedding model for jointly exploring the functional relationships between text and image features for transferring intermodel and intramodel labels to help annotate the images. The intermodal label transfer can be generalized to zero-shot recognition.

Deep embedding

Most of recent zero-shot recognition models have to rely on the state-of-the-art deep convolutional models to extract the image features. As one of the first works, DeViSE [46] extended the deep architecture to learn the visual and semantic embedding,

and it can identify visual objects using both labeled image data as well as semantic information gleaned from unannotated text. ConSE [43] constructed the image embedding approach by mapping images into the semantic embedding space via convex combination of the class label embedding vectors. Both DeViSE and ConSE are evaluated on large-scale data sets—ImageNet [the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)] 2012 1K and ImageNet 2011 21K.

To combine the visual and textual branches in the deep embedding, different loss functions can be considered, including margin-based losses [46] or Euclidean distance loss [53]. Zhang et al. [76] employed the visual space as the embedding space and proposed an end-to-end deep-learning architecture for zero-shot recognition. Their networks have two branches: a visual encoding branch, which uses CNNs to encode the input image as a feature vector, and the semantic embedding branch, which encodes the input semantic representation vector of each class to which the corresponding image belongs.

Recognition models in the embedding space

Once the embedding model is learned, the testing instances can be projected into this embedding space. The recognition can be carried out by using different recognition models. The most commonly used one is the nearest neighbor classifier, which classifies the testing instances by assigning the class label in terms of the nearest distances of the class prototypes against the projections of testing instances in the embedding space. Fu et al. [7] proposed a semilatent zero-shot learning algorithm to update the class prototypes by one-step self-training.

Manifold information can be used in the recognition models in the embedding space. Fu et al. [77] proposed a hypergraph structure in their multiview embedding space; and zero-shot recognition can be addressed by label propagation from unseen prototype instances to unseen testing instances. Changpinyo et al. [78] synthesized classifiers in the embedding space for zero-shot recognition. For multilabel zero-shot learning, the recognition models have to consider the co-occurrence/correlations of different semantic labels [41], [45], [79].

Latent SVM structures have also been used as the recognition models [12], [80]. Wang et al. [80] treated the object attributes as latent variables and learned the correlations of attributes through an undirected graphical model. Hwang et al. [12] utilized a kernelized multitask feature-learning framework to learn the sharing features between objects and their attributes. Additionally, Long et al. [81] employed the attributes to synthesize unseen visual features at the training stage and, thus, zero-shot recognition can be solved by the conventional supervised classification models.

Problems in zero-shot recognition

There are two intrinsic problems in zero-shot recognition—projection domain shift and hubness.

Projection domain shift problems

The projection domain shift problem in zero-shot recognition was first identified by Fu et al. [77]. This problem can be explained as follows: since the source and target data sets have different classes, the underlying data distribution of these classes may also differ. The projection functions learned on the source data set, from visual space to the embedding space, without any adaptation to the target data set, will cause an unknown shift/bias. Figure 1 from [77] gives a more intuitive illustration of this problem. It plots the 85-dimensional (85-D) attribute space representation spanned by feature projections that are learned from source data, and class

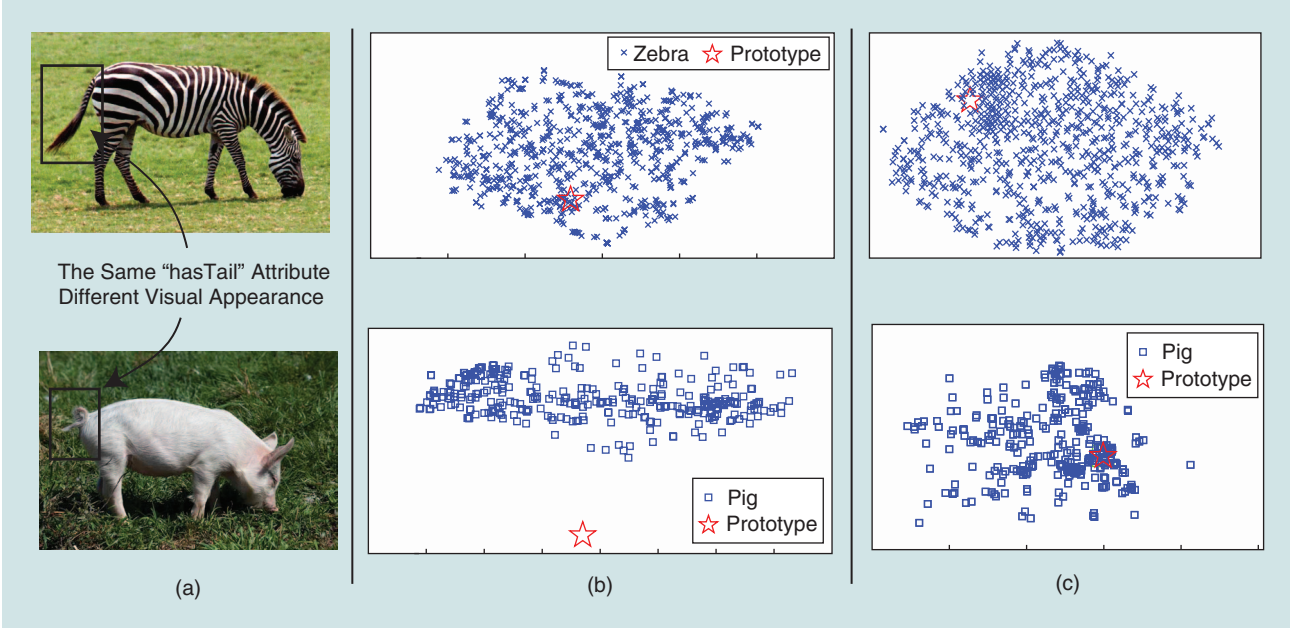

FIGURE 1. Illustrating the projection domain shift problem. (a) The visual space, (b) attribute space, and (c) multiview embedding space. Zero-shot prototypes are annotated as red stars and predicted semantic attribute projections are shown in blue. Both pig and zebra share the same "hasTail" attribute yet with a very different visual appearance of a tail. (Figure used with permission from [77].)

prototypes, which are 85-D binary attribute vectors. Zebra and pig are one of the auxiliary and target classes, respectively; and the same "hasTail" semantic attribute means very different visual appearances for the pig and the zebra. In the attribute space, directly using the projection functions learned from source data sets (e.g., zebra) on the target data sets (e.g., pig) will lead to a large discrepancy between the class prototype of the target class and the predicted semantic attribute projections.

To alleviate this problem, the transductive learning-based approaches were proposed to utilize the manifold information of the instances from unseen classes [69], [77], [82]–[84]. Nevertheless, the transductive setting assumes that all of the testing data can be accessed at once, which obviously is invalid if the new unseen classes appear dynamically and unavailable before learning models. Thus, inductive learning-based approaches [59], [71], [78], [82], [84] have also been studied, and these methods usually enforce other additional constraints or information from the training data.

Hubness problem

The hubness problem is another interesting phenomenon that may be observed in zero-shot recognition. Essentially, the hubness problem can be described as the presence of "universal" neighbors, or hubs, in the space. Radovanovic et al. [85] was the first to study the hubness problem; in [85], a hypothesis is made that hubness is an inherent property of data distributions in the high-dimensional vector space. Nevertheless, Low et al. [86] challenged this hypothesis and showed the evidence that hubness is rather a boundary effect or, more generally, an effect of a density gradient in the process of data generation. Interestingly, their experiments showed that the hubness phenomenon can also occur in low-dimensional data.

While causes for hubness are still under investigation, recent works reported that the regression-based zero-shot learning methods do suffer from this problem. To alleviate this problem, Dinu et al. [87] utilized the global distribution of feature instances of unseen data, i.e., in a transductive manner. In contrast, Yutaro et al. [88] addressed this problem in an inductive way by embedding the class prototypes into a visual feature space.

Beyond zero-shot recognition

Generalized zero-shot recognition and open-set recognition

In conventional supervised learning tasks, it is taken for granted that the algorithms should take the form of "closed set," where all testing classes should be known at training time. Zero-shot recognition, in contrast, assumes that the source and target classes cannot be mixed and that the testing data come from only the unseen classes. This assumption greatly and unrealistically simplifies the recognition tasks. To relax the settings of zero-shot recognition and investigate recognition tasks in a more generic setting, there are several tasks advocated beyond the conventional zero-shot recognition. In particular, generalized zero-shot recognition [89] and open-set recognition tasks have been discussed recently [90]–[92].

The generalized zero-shot recognition proposed in [89] broke the restricted nature of conventional zero-shot recognition and included the training classes among the testing data. Chao et al. [89] showed that it is nontrivial and ineffective to directly extend the current zero-shot learning approaches to solve the generalized zero-shot recognition. Such a generalized setting, due to the more practical nature, is recommended as the evaluation settings for zero-shot recognition tasks [93].

Open-set recognition, in contrast, has been developed independently of zero-shot recognition. Initially, open-set recognition aimed to break the limitation of the closed-set recognition setup. Specifically, the task of open-set recognition tries to identify the class name of an image from a very large set of classes, which includes but is not limited to training classes. The open-set recognition can be roughly divided into two subgroups: conventional open-set recognition and generalized open-set recognition.

Conventional open-set recognition

First formulated in [90], the conventional open-set recognition only identifies whether the testing images come from the training classes or some unseen classes. This category of methods does not explicitly predict to which unseen classes the testing instance (out of seen classes) belongs. In such a setting, the conventional open-set recognition is also known as *incremental learning* [94].

Generalized open-set recognition

The key difference of the conventional open-set recognition is that the generalized open-set recognition also needs to explicitly predict the semantic meaning (class) of testing instances, even from the unseen novel classes. This task was first defined and evaluated in [91] and [92] on the tasks of object categorization. The generalized open-set recognition can be taken as a general version of zero-shot recognition, where the classifiers are trained from training instances of limited training classes, while the learned classifiers are required to classify the testing instances from a very large set of open vocabulary, say, 310,000 class vocabulary in [91] and [92]. Conceptually similar, there are vast variants of generalized open-set recognition tasks that have been studied in other research communities such as open-world person reidentification [95] or open vocabulary scene parsing [96].

One-shot recognition

A closely related problem to zero-shot learning is the one-shot or few-shot learning problem—instead of having only textual description of the new classes, one-shot learning assumes that there are one or few training samples for each class. Similar to zero-shot recognition, one-shot recognition is inspired by humans' ability to learn new object categories from one or very few examples [97]. Existing one-shot learning approaches can be divided into two groups: the direct supervised learning-based approaches and the transfer learning-based approaches.

Direct supervised learning-based approaches

Early approaches do not assume that there is a set of auxiliary classes that are related and/or have ample training samples

Table 2. Data sets in zero-shot recognition. The data sets are divided into three groups: general image classification (A), fine-grained image classification (B), and video classification data sets (C).

	Data Set	Number of Instances	Number of Classes	Number of Attributes	Annotation Level
A	AwA	30,475	50	85	Per class
	aPascal-aYahoo	15,339	32	64	Per image
	PubFig	58,797	200	—	Per image
	PubFig-sub	772	Eight	11	Per image pairs
	OSR	2,688	Eight	6	Per image pairs
	ImageNet	15 million	22,000	—	Per image
	ILSVRC 2010	1.2 million	1,000	—	Per image
	ILSVRC 2012	1.2 million	1,000	—	Per image
B	Oxford 102 Flower	8,189	102	—	—
	CUB-200-2011	11,788	200	312	Per class
	SUN-attribute	14,340	717	102	Per image
C	USAA	1,600	Eight	69	Per video
	UCF101	13,320	101	—	Per video
	ActivityNet	27,801	203	—	Per video
	Fudan-Columbia video (FCVID)	91,223	239	—	Per video

whereby transferable knowledge can be extracted to compensate for the lack of training samples. Instead, the target classes are used to train a standard classifier using supervised learning. The simplest method is to employ nonparametric models such as k-nearest neighbor (kNN), which are not restricted by the number of training samples. However, without any learning, the distance metric used for kNN is often inaccurate. To overcome this problem, metric embedding can be learned and then used for kNN classification. Other approaches attempt to synthesize more training samples to augment the small training data set [97]. However, without knowledge transfer from other classes, the performance of direct supervised learning-based approaches is typically weak. These models cannot meet the requirement of lifelong learning, i.e., when new unseen classes are added, the learned classifier should still be able to recognize the seen existing classes.

Transfer learning-based one-shot recognition

This category of approaches follow a similar setting to zero-shot learning, that is, they assume that an auxiliary set of training data from different classes exist. They explore the paradigm of learning to learn [9] or metalearning [98] and aim to transfer knowledge from the auxiliary data set to the target data set with one or few examples per class. These approaches differ in 1) what knowledge is transferred and 2) how the knowledge is represented. Specifically, the knowledge can be extracted and shared in the form of model prior in a generative model [99], features [100], and semantic attributes [4], [7], [83]. Many of these approaches take a similar strategy as the existing zero-shot learning approaches and transfer knowledge via a shared embedding space. Embedding space can typically be formulated using neural networks (e.g., the Siamese network [101]), discriminative classifiers (e.g., support vector regressors (SVRs) [4], [10]), or kernel embedding [100] methods. Particularly, one of most common methods of embedding is semantic embedding, which is normally explored by projecting the visual features and semantic entities into a common new space. Such projections can take various forms with corresponding loss functions, such as SJE [67], WSABIE [102], ALE [60], DeViSE [46], and CCA [69].

More recently, deep metalearning has received increasing attention for few-shot learning [33], [95], [97], [101], [103]–[105]. Wang et al. [106] proposed the idea of one-shot adaptation by automatically learning a generic, category agnostic transformation from models learned from few samples to models learned from large-enough sample sets. A model-agnostic metalearning framework is proposed by Finn et al. [107], which trains a deep model from the auxiliary data set with the objective that the learned model can be effectively updated/fine-tuned on the new classes with one or few gradient steps. Note that, similar to the generalized zero-shot learning setting, the problem of adding new classes to a deep neural network while keeping the ability to recognize the old classes recently has been attempted [108]. However, the problem of lifelong learning and progressively adding new classes with few-shot remains an unsolved problem.

Data sets in zero-shot recognition

This section summarizes the data sets used for zero-shot recognition. Recently, with the increasing number of proposed zero-shot recognition algorithms, Xian et al. [93] compared and analyzed a significant number of the state-of-the-art methods in depth, and they defined a new benchmark by unifying both the evaluation protocols and data splits. The details of these data sets are listed in Table 2.

Standard data sets

Animals with attributes data set

The Animals with Attributes (AwA) data set [4] consists of the 50 Osher-son/Kemp animal category images collected online. There are 30,475 images with at least 92 examples of each class. Seven different feature types are provided: RGB color histograms, scale-invariant feature transform (SIFT), rgSIFT, pyramid histogram of oriented gradients, speeded up robust features, local self-similarity histograms, and DeCaf. The AwA data set defines 50 classes of animals, and 85 associated attributes (such as "furry" and "has claws"). For the consistent evaluation of attribute-based object classification methods, the AwA data set defined ten test classes: chimpanzee, giant panda, hippopotamus, humpback whale,

leopard, pig, raccoon, rat, Persian cat, and seal. The 6,180 images of those classes are taken as the test data, whereas the 24,295 images of the remaining 40 classes can be used for training. Since the images in AwA are not available under a public license, Xian et al. [93] introduced another new zero-shot learning data set, AWA2, which includes 37,322 publicly licensed and released images from the same 50 classes and 85 attributes as AwA.

The aPascal-aYahoo data set

The aPascal-aYahoo data set [10] has a 12,695-image subset of the PASCAL VOC 2008 data set with 20 object classes (aPascal); and 2,644 images that were collected using the Yahoo image search engine (aYahoo) of 12 object classes. Each image in this data set has been annotated with 64 binary attributes that characterize the visible objects.

CUB-200-2011 data set

CUB-200-2011 [109] contains 11,788 images of 200 bird classes. This is a more challenging data set than AwA—it is designed for fine-grained recognition and has more classes but fewer images. All images are annotated with bounding boxes, part locations, and attribute labels. Images and annotations were filtered by multiple users of AMT. CUB-200-2011 is used as the benchmark data set for multiclass categorization and part localization. Each class is annotated with 312 binary attributes derived from the bird species ontology. A typical setting is to use 150 classes as auxiliary data, holding out 50 as target data, which is the setting adopted in Akata et al. [60].

The outdoor scene recognition data set

The outdoor scene recognition (OSR) [110] data set consists of 2,688 images from eight categories and six attributes (openness, natural, etc.) and an average 426 labeled pairs for each attribute from 240 training images. Graphs constructed are thus extremely sparse. Pairwise attribute annotation was collected by AMT (Kovashka et al. [25]). Each pair was labeled by five workers to average the comparisons by majority voting. Each image also belongs to a scene type.

Public figure face database

The public figure face (PubFig) database [3] is a large face data set of 58,797 images of 200 people collected from the Internet. Parikh et al. [13] selected a subset of PubFig consisting of 772 images from eight people and 11 attributes (including "smiling," "round face," etc.). We annotate this subset as PubFig-sub. The pairwise attribute annotation was collected by AMT [25]. Each pair was labeled by five workers. A total of 241 training images for PubFig-sub were labeled. The average number of compared pairs per attribute was 418.

SUN attribute data set

The SUN attribute data set [111] is a subset of the SUN database [112] for fine-grained scene categorization, and it has 14,340 images from 717 classes (20 images per class). Each image is annotated with 102 binary attributes that describe the scenes' material and surface properties as well as lighting conditions, functions, affordances, and general image layout.

Unstructured social activity attribute data set

The unstructured social activity attribute (USAA) data set [5] is the first benchmark video attribute data set for social activity video classification and annotation. The ground-truth attributes are annotated for eight semantic class videos of the Columbia Consumer Video data set [113] and select 100 videos per-class for training and testing, respectively. These classes were selected as the most complex social group activities. By referring to the existing work on video ontology [113], the 69 attributes can be divided into five broad classes: actions, objects, scenes, sounds, and camera movement. Directly using the ground-truth attributes as input to an SVM, the videos can have 86.9% classification accuracy. This illustrates the challenge of the USAA data set: while the attributes are informative, there is sufficient intraclass variability in the attribute-space, and even perfect knowledge of the instance-level attributes is also insufficient for perfect classification.

ImageNet data sets

ImageNet has been used in several different papers with relatively different settings. The original ImageNet data set has been proposed in [114]. The full set of ImageNet contains over 15 million labeled high-resolution images belonging to roughly 22,000 categories and labeled by human annotators using the AMT crowdsourcing tool. Started in 2010 as part of the Pascal Visual Object Challenge, the annual competition ILSVRC has been held. ILSVRC uses a subset of ImageNet with roughly 1,000 images in each of 1,000 categories. In [40] and [83], Robhrbach et al. split the ILSVRC 2010 data into 800 classes for source data and 200 classes for target data. In [91], Fu et al. employed the training data of ILSVRC 2012 as the source data, and the testing part of ILSVRC 2012 as well as the data of ILSVRC 2010 as the target data. The full-sized ImageNet data has been used in [43], [46], and [78].

Oxford 102 flower data set

The Oxford 102 flower data set [115] is a collection of 102 groups of flowers, each with 40–256 flower images and a total of 8,189 images. The flowers were chosen from the common flower species in the United Kingdom. Elhoseiny et al. [52] generated textual descriptions for each class of this data set.

UCF101 data set

The UCF101 data set [116] is another popular benchmark for human action recognition in videos, which consists of 13,320 video clips (27 h in total) with 101 annotated classes. More recently, the THUMOS-2014 Action Recognition Challenge [117] created a benchmark by extending upon the UCF-101 data set (used as the training set). Additional videos were collected from the Internet, including 2, 500 background videos, 1,000 validation and 1,574 test videos.

FCVID data set

The FCVID data set [118] contains 91,223 web videos annotated manually into 239 categories. Categories cover a wide range of topics (not only activities), such as social events (e.g., tailgate party), procedural events (e.g., making a cake), object appearances

(e.g., panda), and scenic videos (e.g., beach). A standard split consists of 45,611 videos for training and 45,612 videos for testing.

ActivityNet data set

Released in 2015, ActivityNet [119] is another large-scale video data set for human activity recognition and understanding. It consists of 27,801 video clips annotated into 203 activity classes, totaling 849 h of video. Compared with existing data sets, ActivityNet has more fine-grained action categories (e.g., drinking beer and drinking coffee). ActivityNet had the settings of both trimmed and untrimmed videos of its classes.

Discussion of data sets

In Table 2, we roughly divide all the data sets into three groups: general image classification, fine-grained image classification, and video classification. These data sets have been employed widely as the benchmark in many previous works. However, we believe that when making a comparison with the other existing methods on these data sets, there are several issues that should be discussed.

Features

With the renaissance of deep CNNs, deep features of images/videos have been used for zero-shot recognition. Note that different types of deep features (e.g., Overfeat, VGG-19, or ResNet) have varying levels of semantic abstraction and representation ability; and even the same type of deep features, if fine-tuned on different data sets and with slightly different parameters, will also have different representative ability. Thus, without using the same type of features, it is not possible to conduct a fair comparison among different methods and draw any meaningful conclusion. It is important to note that it is possible that the improved performance of one zero-shot recognition could be largely attributed to the better deep features used.

Auxiliary data

As mentioned, zero-shot recognition can be formulated in a transfer learning setting. The size and quality of auxiliary data can be very important for the overall performance of zero-shot recognition. Note that these auxiliary data do not only include the auxiliary source image/video data set, but also refer to the data to extract/train the concept ontology, or semantic word vectors. For example, the semantic word vectors trained on large-scale linguistic articles, in general, are better semantically distributed than those trained on small-sized linguistic corpus. Similarly, GloVe [120] is reported to be better than the skip-gram and CBOW models [55]. Therefore, to make a fair comparison with existing works, another important factor is to use the same set of auxiliary data.

Evaluation

For many data sets, there is no agreed source/target splits for zero-shot evaluation. Xian et al. [93] suggested a new benchmark by unifying both the evaluation protocols and data splits.

Future research directions

More generalized and realistic setting

A detailed review of existing zero-shot learning methods shows that, overall, the existing efforts have been focused on a rather restrictive and impractical setting: classification is required for new object classes only, and the new unseen classes, although having no training sample present, are assumed to be known. In reality, one wants to progressively add new classes to the existing classes. This needs to be achieved without jeopardizing the ability of the model to recognize existing seen classes. Furthermore, we cannot assume that the new samples will only come from a set of known unseen classes. Rather, they can only be assumed to belong to either existing seen classes, known unseen classes, or unknown unseen classes. We therefore foresee that a more generalized setting will be adopted by future zero-shot learning works.

Combining zero-shot with few-shot learning

As mentioned previously, the problems of zero-shot and few-shot learning are closely related and, as a result, many existing methods use the same or similar models. However, it is somewhat surprising to note that no serious efforts have been taken to address these two problems jointly. In particular, zero-shot learning would typically not consider the possibility of having few training samples, while few-shot learning ignores the fact that the textual description/human knowledge about the new class is always there to be exploited. A few existing zero-shot learning methods [7], [54], [91] have included few-shot learning experiments. However, they typically use a naive k-NN approach, i.e., each class prototype is treated as a training sample and together with the k-shot, this becomes a k+1-shot recognition problem. However, as shown by existing zero-shot learning methods [77], the prototype is worth far more that one training sample; thus, it should be treated differently. We then expect a future direction on extending the existing few-shot learning methods by incorporating the prototype as a "supershot" to improve the model learning.

Beyond object categories

So far, current zero-shot learning efforts are limited to recognizing object categories. However, visual concepts can have far more complicated relationships than object categories. In particular beyond objects/nouns, attributes/adjectives are important visual concepts. When combined with objects, the same attribute often has different meaning, e.g., the concept of *yellow* in a yellow face and a yellow banana clearly differs. Zero-shot learning attributes with associated objects is thus an interesting future research direction.

Curriculum learning

In a lifelong learning setting, a model will incrementally learn to recognize new classes while keeping the capacity for existing classes. A related problem is thus how to select the more suitable new classes to learn given the existing classes. It has been shown that the sequence of adding different classes has a clear impact on the model performance [94]. It is therefore useful to investigate how to incorporate the curriculum learning principles in designing a zero-shot learning strategy.

Conclusions

In this article, we have reviewed the recent advances in zero-shot recognition. First, different types of semantic representations are examined and compared; the models used in zero-shot learning have also been investigated. Next, beyond zero-shot recognition, one-shot and open-set recognition are identified as two very important related topics and thus reviewed. Finally, the commonly used data sets in zero-shot recognition have been reviewed with a number of issues in existing evaluations of zero-shot recognition methods discussed. We also point out a number of research direction that we believe will be the focus of future zero-shot recognition studies.

Acknowledgments

This work is supported in part by three grants from National Science Foundation China (number 61702108, number 61622204, and number 61572134), and a European FP7 project (PIRSESGA-2013–612652). Yanwei Fu is supported by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (number TP2017006). Prof. Yu-Gang Jiang is the corresponding author of this article.

Authors

Yanwei Fu (yanweifu@fudan.edu.cn) received his B.Sc. degree in information and computing sciences and his M.Eng. degree in computer science from the Department of Computer Science and Technology at Nanjing University, China, in 2008 and 2011, respectively. He is now pursuing his Ph.D. degree in the Vision Group of the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include attribute learning, topic models, learning to rank, video summarization, and image segmentation.

Tao Xiang (t.xiang@qmul.ac.uk) received his B.Sc. degree in electronic engineering from Xi'an Jiaotong University, China, in 1995 and his Ph.D. degree in electrical and computer engineering from the National University of Singapore in 2002. He is currently a reader (associate professor) in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include computer vision, machine learning, and data mining. He has published more than 140 papers in international journals and conferences.

Yu-Gang Jiang (ygj@fudan.edu.cn) received his Ph.D. degree in computer science from City University of Hong Kong in 2009. He is a professor with the School of Computer Science, Fudan University, China. His Lab for Big Video Data Analytics conducts research on all aspects of extracting high-level information from big video data, such as video event recognition, object/scene recognition and large-scale visual search. His work has led to many awards, including the Rising Star Awards from the China Council of Association for Computing Machinery (ACM) and the Special Interest Group on Multimedia of ACM.

Xiangyang Xue (xyxue@fudan.edu.cn) received his B.S., M.S., and Ph.D. degrees in communication engineering from Xidian University, Xi'an, China, in 1989, 1992, and 1995, respectively. He is currently a computer science professor at Fudan University, Shanghai, China. His research interests include multimedia information processing and machine learning.

Leonid Sigal (lsigal@cs.ubc.ca) received his M.Sc. degree from Brown University, Providence, Rhode Island in 2003, his M.A. degree from Boston University in 1999, and his Ph.D. degree from Brown University in 2008, all in computer science. He is an associate professor at the University of British Columbia, Canada. Prior to this, he was a senior research scientist at Disney Research. His research interests lie in the areas of computer vision, machine learning, and computer graphics, with emphasis on machine learning and statistical approaches for visual recognition, understanding, and analytics. He has had more than 70 papers published in these fields.

Shaogang Gong (s.gong@qmul.ac.uk) received his B.Sc. degree in information engineering in 1985 from the University of Electronic Science and Technology of China and his D.Phil. degree in computer vision in 1989 from the University of Oxford. He has been a professor of visual computation at Queen Mary University of London since 2001. He is a fellow of the Institution of Electrical Engineers and the British Computer Society. His research interests include computer vision, machine learning, and video analysis.

References

[1] S. Thrun and T. M. Mitchell, "Lifelong robot learning," *Robot. Autom. Syst.*, vol. 15, no. 1–2, pp. 25–46, 1995.

[2] M. Palatucci, G. Hinton, D. Pomerleau, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Proc. Neural Information Processing Systems Conf.*, 2009, pp. 1410–1418.

[3] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE Int. Conf. Computer Vision*, 2009, pp. 365–372.

[4] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, 2013.

[5] Y. Fu, T. Hospedales, T. Xiang, and S. Gong, "Attribute learning for understanding unstructured social activity," in *Proc. European Conf. Computer Vision*, 2012. pp. 530–543.

[6] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 3337–3344.

[7] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Learning multi-modal latent attributes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 36, no. 2, pp. 303–316, 2013.

[8] R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. D. Manning, and A. Y. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. Neural Information Processing Systems Conf.*, 2013, pp. 935–943.

[9] S. Thrun, *Learning to Learn: Introduction*. Norwell, MA: Kluwer Academic Publishers, 1996.

[10] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 1778–1785.

[11] X. Wang and T. Zhang. (2011). Clothes search in consumer photos via color matching and attribute learning, in *Proc. ACM Int. Conf. Multimedia*. [Online]. Available: http://doi.acm.org/10.1145/2072298.2072013

[12] S. J. Hwang, F. Sha, and K. Grauman, "Sharing features between objects and their attributes," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 1761–1768.

[13] D. Parikh and K. Grauman, "Relative attributes," in *Proc. IEEE Int. Conf. Computer Vision,* 2011, pp. 503–510.

[14] Y. Fu, T. M. Hospedales, J. Xiong, T. Xiang, S. Gong, Y. Yao, and Y. Wang, "Robust estimation of subjective visual properties from crowdsourced pairwise labels," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 38, no. 3, pp. 563–577 2016.

[15] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. V. Gool, "The interestingness of images," in *Proc. IEEE Int. Conf. Computer Vision*, 2013, pp. 1633–1640.

[16] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 2467–2474.

[17] Y.-G. Jiang, YanranWang, R. Feng, X. Xue, Y. Zheng, and H. Yang, "Understanding and predicting interestingness of videos," in *Proc. Conf. Association Advancement Artificial Intelligence*, 2013, pp. 1113–1119.

[18] E. M. Rudd, M. Gunther, and T. E. Boult, "MOON: A mixed objective optimization network for the recognition of facial attributes," in *Proc. European Conf. Computer Vision*, 2016, pp. 19–35.

[19] J. Wang, Y. Cheng, and R. S. Feris, "Walk and learn: Facial attribute representation learning from egocentric video and contextual data," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 2295–2304.

[20] A. Datta, R. Feris, and D. Vaquero, "Hierarchical ranking of facial attributes," in *Proc. IEEE Int. Conf. Automatic Face Gesture Recognition*, 2011, pp. 36–42.

[21] Z. Wang, K. He, Y. Fu, R. Feng, Y.-G. Jiang, and X. Xue, "Multi-task deep neural network for joint face recognition and facial attribute prediction," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2017, pp. 365–374.

[22] D. F. Fouhey, A. Gupta, and A. Zisserman, "Understanding higher-order shape via 3D shape attributes," *IEEE Trans. Pattern Anal. Machine Intell.*, 2017.

[23] G. Wang and D. Forsyth, "Joint learning of visual attributes, object classes and visual saliency," in *Proc. IEEE Int. Conf. Computer Vision*, 2009, pp. 537–544.

[24] V. Ferrari and A. Zisserman, "Learning visual attributes," in *Proc. Neural Information Processing Systems Conf.*, Dec. 2007, pp. 433–440.

[25] A. Kovashka, D. Parikh, and K. Grauman, "WhittleSearch: Image search with relative attribute feedback," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 2973–2980.

[26] A. Shrivastava, S. Singh, and A. Gupta, "Constrained semi-supervised learning via attributes and comparative attributes," in *Proc. European Conf. Computer Vision*, 2012, pp. 369–383.

[27] A. Biswas and D. Parikh, "Simultaneous active learning of classifiers and attributes via relative feedback," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 644–651.

[28] A. Parkash and D. Parikh, "Attributes for classifier feedback," in *Proc. European Conf. Computer Vision*, 2012, pp. 354–368.

[29] K. K. Singh and Y. J. Lee, "End-to-end localization and ranking for relative attributes," in *Proc. European Conf. Computer Vision*, 2016, pp. 753–769.

[30] E. Rudd, M. Günther, and T. Boult, "MOON: A mixed objective optimization network for the recognition of facial attributes," in *Proc. European Conf. Computer Vision*, 2016, pp. 19–35.

[31] D. Parikh and K. Grauman, "Interactively building a discriminative vocabulary of nameable attributes," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 1681–1688.

[32] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. (2009). Inferring semantic concepts from community-contributed images and noisy tags, *Proc. ACM Int. Conf. Multimedia*. [Online]. Available: http://doi.acm.org/10.1145/1631272.1631305

[33] A. Habibian, T. Mensink, and C. Snoek, "Videostory: A new multimedia embedding for few-example recognition and translation of events," in *Proc. ACM Multimedia Conf.*, 2014, pp. 17–26.

[34] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, "Adding semantics to detectors for video retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 975–986, 2007.

[35] G. Toderici, H. Aradhye, M. Pasca, L. Sbaiz, and J. Yagnik, "Finding meaning on youtube: Tag recommendation and category discovery," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010, pp. 3447–3454.

[36] Z. Wu, Y. Fu, Y.-G. Jiang, and L. Sigal, "Harnessing object and scene semantics for large-scale video understanding," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 3112–3121.

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

[38] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. (2007). Correlative multi-label video annotation, *Proc. ACM Int. Conf. Multimedia*. [Online]. Available: http://doi.acm.org/10.1145/1291233.1291245

[39] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba, "Semantic label sharing for learning with many categories," in *Proc. European Conf. Computer Vision*, 2010, 762–775.

[40] M. Rohrbach, M. Stark, and B. Schiele, "Evaluating knowledge transfer and zero-shot learning in a large-scale setting," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 1641–1648.

[41] T. Mensink, E. Gavves, and C. G. Snoek, "Costa: Co-occurrence statistics for zero-shot classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 2441–2448.

[42] C. Gan, Y. Yang, L. Zhu, and Y. Zhuang, "Recognizing an action using its name: A knowledge-based approach," *Int. J. Comput. Vis.*, vol. 120, no. 1, pp. 61–77, 2016.

[43] 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. Learning Representations*, 2014.

[44] Z. Zhang and V. Saligrama, "Zero-shot recognition via structured prediction," in *Proc. European Conf. Computer Vision*, 2016, pp. 533–548.

[45] Y. Fu, Y. Yang, T. Hospedales, T. Xiang, and S. Gong, "Transductive multi-label zero-shot learning," in *Proc. British Machine Vision Conf.*, 2014.

[46] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "DeViSE: A deep visual-semantic embedding model," in *Proc. Neural Information Processing Systems Conf.*, 2013, pp. 2121–2129.

[47] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, "Improving word representations via global context and multiple word prototypes," in *Proc. Association Computational Linguistics Conf.*, 2012, pp. 873–882.

[48] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization," *IEEE Trans. Affect. Comput.*, 2016.

[49] X. Chang, Y. Yang, G. Long, C. Zhang, and A. Hauptmann, "Dynamic concept composition for zero-example event detection," in *Proc. Conf. Association Advancement Artificial Intelligence*, 2016, pp. 3464–3470.

[50] J. Qin, L. Liu, L. Shao, F. Shen, B. Ni, J. Chen, and Y. Wang, "Zero-shot action recognition with error-correcting output codes," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 2833–2842.

[51] S. Wu, F. Luisier, and S. Bondugula, "Zero-shot event detection using multimodal fusion of weakly supervised concepts," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 2665–2672.

[52] M. Elhoseiny, B. Saleh, and A. Elgammal, "Write a classifier: Zero-shot learning using purely textual descriptions," in *Proc. IEEE Int. Conf. Computer Vision*, Dec. 2013, pp. 2584–2591.

[53] J. L. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov, "Predicting deep zero-shot convolutional neural networks using textual descriptions," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 4247–4255.

[54] S. Reed, Z. Akata, B. Schiele, and H. Lee, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 49–58.

[55] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Neural Information Processing Systems Conf.*, 2013, pp. 3111–3119.

[56] Z. S. Harris, *Distributional Structure*. Dordrecht, The Netherlands: Springer, 1981, pp. 3–22.

[57] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks," in *Proc. Association Advancement Artificial Intelligence Conf.*, 2008, pp. 646–651.

[58] X. Yu and Y. Aloimonos, "Attribute-based transfer learning for object categorization with zero/one training example," in *Proc. European Conf. Computer Vision*, 2010, pp. 127–140.

[59] D. Jayaraman and K. Grauman, "Zero shot recognition with unreliable attributes," in *Proc. Neural Information Processing Systems Conf.*, 2014, pp. 3464–3472.

[60] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 819–826.

[61] J. Weston, S. Bengio, and N. Usunier, "Large scale image annotation: Learning to rank with joint word-image embeddings," *Machine Learning*, vol. 81, no. 1, pp. 21–35, 2010.

[62] Z. Li, E. Gavves, T. E. J. Mensink, and C. G. M. Snoek, "Attributes make sense on segmented objects," in *Proc. European Conf. Computer Vision*, 2014, pp. 350–365.

[63] X. Li, Y. Guo, and D. Schuurmans, "Semi-supervised zero-shot classification with label representation learning," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 4211–4219.

[64] Z. Zhang and V. Saligrama, "Zero-shot learning via joint latent similarity embedding," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 6034–6042.

[65] D. Jayaraman, F. Sha, and K. Grauman, "Decorrelating semantic visual attributes by resisting the urge to share," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1629–1636.

[66] S. J. Hwang and L. Sigal, "A unified semantic embedding: Relating taxonomies and attributes," in *Proc. Neural Information Processing Systems Conf.*, 2014, pp. 271–279.

[67] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 2927–2936.

[68] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *Proc. Int. Conf. Machine Learning*, 2015, pp. 2152–2161.

[69] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong, "Transductive multi-view embedding for zero-shot recognition and annotation," in *Proc. European Conf. Computer Vision*, 2014, pp. 584–599.

[70] D. Mahajan, S. Sellamanickam, and V. Nair, "A joint learning framework for attribute models and object descriptions," in *Proc. IEEE Int. Conf. Computer Vision*, 2011, pp. 1227–1234.

[71] Z. Fu, T. Xiang, E. Kodirov, and S. Gong, "Zero-shot object recognition by semantic manifold distance," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 2635–2644.

[72] 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. Computer Vision*, 2014, pp. 48–64.

[73] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[74] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *Int. J. Computer Vision*, pp. 210–233, 2013.

[75] G.-J. Qi, W. Liu, C. Aggarwal, and T. Huang, "Joint intermodal and intramodal label transfers for extremely rare or unseen classes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 39, no. 7, pp. 1360–1373, 2017.

[76] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[77] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 37, no. 11, pp. 2332–2345, 2015.

[78] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 5327–5336.

[79] Y. Zhang, B. Gong, and M. Shah, "Fast zero-shot image tagging," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 5985–5994.

[80] Y. Wang and G. Mori, "A discriminative latent model of image region and object tag correspondence," in *Proc. Neural Information Processing Systems Conf.*, 2010, pp. 2397–2405.

[81] Y. Long, L. Liu, L. Shao, F. Shen, G. Ding, and J. Han, "From zero-shot learning to conventional supervised classification: Unseen visual data synthesis," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[82] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Unsupervised domain adaptation for zero-shot learning," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 2452–2460.

[83] M. Rohrbach, S. Ebert, and B. Schiele, "Transfer learning in a transductive setting," in *Proc. Neural Information Processing Systems Conf.*, 2013, pp. 46–54.

[84] Y. Li, D. Wang, H. Hu, Y. Lin, and Y. Zhuang, "Zero-shot recognition using dual visual-semantic mapping paths," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[85] B. Marco, L. Angeliki, and D. Georgiana, "Hubness and pollution: Delving into cross-space mapping for zero-shot learning," in *Proc. Association Computational Linguistics Conf.*, 2015, pp. 270–280.

[86] T. Low, C. Borgelt, S. Stober, and A. Nürnberger, *The Hubness Phenomenon: Fact or Artifact?* Springer: Berlin, Heidelberg, 2013.

[87] G. Dinu, A. Lazaridou, and M. Baroni, "Improving zero-shot learning by mitigating the hubness problem," in *Proc. Int. Conf. Learning Representations Workshop*, 2014, pp. 267–278.

[88] Y. Shigeto, I. Suzuki, K. Hara, M. Shimbo, and Y. Matsumoto, "Ridge regression, hubness, and zero-shot learning," in *Proc. European Conf. Machine Learning Principles Practice Knowledge Discovery Databases*, 2015, pp. 135–151.

[89] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha., "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *Proc. European Conf. Computer Vision*, 2016, pp. 52–68.

[90] W. J. Scheirer, A. Rocha, A. Sapkota, and T. E. Boult, "Towards open set recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, 2013, pp. 1757–1772.

[91] Y. Fu and L. Sigal, "Semi-supervised vocabulary-informed learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 5337–5346.

[92] Y. Fu, H. Dong, Y. feng Ma, Z. Zhang, and X. Xue, "Vocabulary-informed extreme value learning," arXiv Preprint, arXiv:1705.09887, 2017.

[93] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning: The good, the bad and the ugly," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[94] S. A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "ICARL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[95] W. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE Trans. Pattern Anal. Machine Intell.*, 2016, pp. 591–606.

[96] H. Zhao, X. Puig, B. Zhou, S. Fidler, and A. Torralba, "Open vocabulary scene parsing," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[97] B. M. Lake and R. Salakhutdinov, "One-shot learning by inverting a compositional causal process," in *Proc. Neural Information Processing Systems Conf.*, 2013, pp. 2526–2534.

[98] R. J. Vilalta and Y. Drissi, "A perspective view and survey of meta-learning," *Artif. Intell. Rev.*, vol. 18, no. 2, pp. 77–95, 2002.

[99] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, no. 4, pp. 594–611, 2006.

[100] T. Hertz, A. Hillel, and D. Weinshall, "Learning a kernel function for classification with small training samples," in *Proc. Int. Conf. Machine Learning*, 2016, pp. 401–408.

[101] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. Int. Conf. Machine Learning Deep Learning Workshop*, 2015.

[102] J. Weston, S. Bengio, and N. Usunier, "WSABIE: Scaling up to large vocabulary image annotation," in *Proc. Int. Conf. Artificial Intelligence*, 2011, pp. 2764–2770.

[103] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *Proc. Neural Information Processing Systems Conf.*, 2016, pp. 523–531.

[104] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Neural Information Processing Systems Conf.*, 2016, pp. 3630–3638.

[105] H. Zhang, K. Dana, and K. Nishino, "Friction from reflectance: Deep reflectance codes for predicting physical surface properties from one-shot in-field reflectance," in *Proc. European Conf. Computer Vision*, 2016, pp. 808–824.

[106] Y. Wang and M. Hebert, "Learning to learn: Model regression networks for easy small sample learning," in *Proc. European Conf. Computer Vision*, 2016, pp. 616–634.

[107] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Machine Learning*, 2017, pp. 1126–1135.

[108] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," arXiv Preprint, arXiv:1606.04671, 2016.

[109] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Data Set," California Inst. Technology, Los Angeles, Tech. Rep. CNS-TR-2011-001, 2011.

[110] A. Oliva and A. Torralba, "Modeling the shape of the scene: Aholistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[111] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 2751–2758.

[112] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010, pp. 3485–3492.

[113] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2011.

[114] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[115] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. Indian Conf. Computer Vision Graphics Image Processing*, 2008, pp. 722–729.

[116] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A data set of 101 human action classes from videos in the wild," arXiv Preprint, arXiv:1212.0402, 2012.

[117] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah, "The thumos challenge on action recognition for videos in the wild," *Comput. Vis. Image Understanding*, vol. 155, pp. 1–23 Feb. 2017.

[118] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.- F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," *IEEE Trans. Pattern Anal. Machine Intell.*, 2017, p. 1.

[119] F. C. H. V. E. B. Ghanem and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 961–970.

[120] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Language Processing*, 2014, pp. 1532–1543.

SP