Incomplete Attribute Learning with Auxiliary Labels

Kongming Liang[1,2,3], **Yuhong Guo**[2], **Hong Chang**[1], **Xilin Chen**[1,3]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]School of Computer Science, Carleton University, Ottawa, Canada
[3]University of Chinese Academy of Sciences, Beijing 100049, China
kongming.liang@vipl.ict.ac.cn, yuhong.guo@carleton.ca, {changhong, xlchen}@ict.ac.cn

Abstract

Visual attribute learning is a fundamental and challenging problem for image understanding. Considering the huge semantic space of attributes, it is economically impossible to annotate all their presence or absence for a natural image via crowdsourcing. In this paper, we tackle the incompleteness nature of visual attributes by introducing auxiliary labels into a novel transductive learning framework. By jointly predicting the attributes from the input images and modeling the relationship of attributes and auxiliary labels, the missing attributes can be recovered effectively. In addition, the proposed model can be solved efficiently in an alternative way by optimizing quadratic programming problems and updating parameters in closed-form solutions. Moreover, we propose and investigate different methods for acquiring auxiliary labels. We conduct experiments on three widely used attribute prediction datasets. The experimental results show that our proposed method can achieve the state-of-the-art performance with access to partially observed attribute annotations.

1 Introduction

Attributes are semantic properties of objects which can be inferred from visual images. Beyond traditional object recognition, attribute learning shows a promising way to natural image understanding as it is able to provide fine-grained descriptions. According to the definitions in previous works [Farhadi *et al.*, 2009; Russakovsky and Fei-Fei, 2010], attributes usually contain rich semantic meanings, including color, shape, texture and object parts. Recent research has verified that attributes can benefit many relevant computer vision tasks such as image retrieval [Kovashka *et al.*, 2012; Liang *et al.*, 2016] and image captioning [Fang *et al.*, 2015; You *et al.*, 2016]. Moreover, attribute learning makes it possible to do *zero-shot classification* [Lampert *et al.*, 2014; Jayaraman and Grauman, 2014] by modeling the correlation between seen and unseen object categories.

Direct attribute prediction methods [Farhadi *et al.*, 2009; Lampert *et al.*, 2014] train a binary classifier to predict each individual attribute. Since the binary attribute classifiers are trained independently, they fail to exploit the correlation information between attributes. By taking each attribute as one subtask, [Jayaraman *et al.*, 2014; Chen *et al.*, 2014] formulate attribute learning in a regularization-based multi-task learning framework. In this way, the correlations between attributes are well incorporated during the learning process. In addition to leveraging the correlation within attributes, the relationship between attributes and their associated object categories can also play a key factor for improving the discriminative ability of attribute classifiers. [Wang and Ji, 2013; Huang *et al.*, 2015] propose to model a high order relationship between attribute and object categories. In this way, they can better recognize the attributes which are hard to predict based only on visual appearances. Moreover, by modeling the attribute classifier in a category-specific way [Liang *et al.*, 2015], different visual attribute manifestations across categories (e.g. attribute "fluffy" varies considerably between dog and towel) can be characterized explicitly. Nevertheless, all the above methods have assumed the training images with complete attribute annotations.

Since multiple attributes may present on a single instance and the space of attributes is almost infinite, exhaustively annotating all the presented attributes seems economically infeasible. The resulting incompleteness of attribute labels can increase the difficulty of attribute prediction to a large extent. Therefore, it is very necessary to build an attribute prediction model that can tackle the incomplete problem. Although the incomplete learning problem has received attention in multi-label learning, there is almost no previous work to investigate the problem for attribute learning. In this paper, we propose to tackle the attribute learning problem with incomplete annotations. Our contributions are in four folds: First, we propose a novel transductive learning model to predict visual attributes, which is able to exploit both labeled and unlabeled images in the learning process. Second, we incorporate high-level auxiliary labels into the transductive learning model via label matrix completion to improve attribute prediction. By enforcing the low rank property on the augmented label matrix, the model can infer the missing attributes from both the observed attributes and the augmented high-level auxiliary labels such as auxiliary labels. Third, we investigate different sources of high-level auxiliary labels, including both the existing object category annotations and the knowledge transferred from auxiliary large scale data sources. Finally, we

conduct experiments on the widely used datasets for attribute learning. Experimental results demonstrate the effectiveness of our proposed method on attribute learning with incomplete annotations.

2 Related Work

A number of previous works have been proposed to tackle incomplete label problem in the literature of multi-label learning. Common ways include taking the missing part of labels as negative labels [Sun *et al.*, 2010] or training on the provided labels [Yu *et al.*, 2014]. [Chen *et al.*, 2013] proposed a fast image tagging algorithm with only incomplete tags for image annotation. It co-regularized both the partially observed tags and image representation to recover the complete tag labels within a joint convex loss function. [Wu *et al.*, 2013] proposed to infer the missing labels through label completion based on visual similarity and label co-occurrence. Moreover, [Wu *et al.*, 2015] proposed to complete the missing labels by further adding semantic hierarchy constraints. They addressed the incomplete multi-label learning problem by using a mixed graph to exploit the label dependencies according to instance similarity, class co-occurrence, and semantic hierarchy simultaneously. [Zhao and Guo, 2015] proposed to solve incomplete multi-label learning in a semi-supervised way by integrating a Laplacian manifold regularization into the learning procedure. However, directly using the above methods for incomplete attribute learning is not effective: Since the visual manifestations for a single attribute vary across different object categories, it is difficult to exploit the correlation between attributes when only considering visual appearance.

Transferring auxiliary labels from external knowledge database is an effective way to boost the original learning task. [Hwang and Sigal, 2014] used the taxonomy tree to jointly embed attributes and super-categories into the same space. [Frome *et al.*, 2013] mapped the object category labels to its corresponding semantic embedding. The embeddings of object category labels are learned from textual data in an unsupervised way. [Lu, 2016] proposed an unsupervised zero-shot learning method to embed large scale object classes by exploiting the outputs of a trained neural network. [Lu *et al.*, 2016] leveraged language priors from semantic word embeddings to improve visual relationship detection task. [Liang *et al.*, 2015] leveraged auxiliary object category labels to model the high order relationship between image, object and attribute. A common semantic space is constructed for embedding the three types of information. Inspired by the above methods, we propose different ways to acquire auxiliary labels which are helpful for incomplete attribute learning.

3 Method

In this work, we consider learning image attribute predictors in the following setting. Assume we have an input data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$, which contains n images, and each image is represented as a d-dimensional feature vector. Without loss of generality, we assume the first n_ℓ images, $\mathbf{X}^\ell$, from $\mathbf{X}$ are labeled training instances and associated with an attribute indicator matrix $\mathbf{Y}^\ell \in \{0,1\}^{L\times n_\ell}$, where 1 indicates the presence of the corresponding attribute among the total L predefined attributes, while assuming the attribute indicator matrix $\mathbf{Y}^u \in \{0,1\}^{L\times n_u}$ for the rest n_u (such that $n = n_\ell + n_u$) images, $\mathbf{X}^u$, is unobserved and needs to be predicted. Thus overall we have a partially observed attribute-based label indicator matrix $\mathbf{Y} = [\mathbf{Y}^\ell, \mathbf{Y}^u]$. Below we present a novel transductive learning method for attribute prediction, which is able to exploit auxiliary labels and can be naturally extended to handle incomplete attribute annotations.

3.1 Attribute Learning with Auxiliary Labels

Though attribute learning can be tackled as a standard label prediction problem, the nature of visual attributes enables the existence of related auxiliary label categories on the same images beyond the attribute labels. For example, the object categories can be a natural set of auxiliary labels that can be useful for attribute label prediction. Such auxiliary labels and the target attribute labels can typically present strong correlation patterns and dependence relationships. We hence propose a novel transductive learning model that not only exploits both labeled images and unlabeled images for attribute prediction, but also integrates auxiliary labels into the learning process. In particular, we assume there is a set of $\hat{L}$ auxiliary labels, and the prediction information on these auxiliary labels for all the images can be encoded into a matrix $\mathbf{Z} \in [0,1]^{\hat{L}\times n}$, and we formulate our transductive learning into the following framework:

$$\min_{\mathbf{W},\mathbf{Y}^u,\mathbf{M}} \quad \mathcal{L}(f(\mathbf{X};\mathbf{W}),\mathbf{Y}) + \frac{\alpha}{2}\left\|\begin{bmatrix}\mathbf{Y}^l,\mathbf{Y}^u\\ \mathbf{Z}\end{bmatrix} - \mathbf{M}\right\|_F^2 + \beta\|\mathbf{M}\|_* + \frac{\gamma}{2}\|\mathbf{W}\|_F^2, \quad (1)$$

where $f(\cdot;\mathbf{W})$ denotes the attribute prediction function with model parameter $\mathbf{W}$, and $\mathcal{L}(\cdot,\cdot)$ denotes the attribute prediction loss function; $\|\cdot\|_F$ denotes the Frobenius norm and $\|\cdot\|_*$ denotes the nuclear norm; α and β are trade-off parameters. The nuclear norm enforces the low-rank property over the M matrix. Together with the second term of the objective function, by pushing the augmented label matrix to be close to a low-rank matrix, we aim to capture the linear correlations between the augmented labels and infer the unobserved labels such as Y^u from the observed ones such as Z. The proposed framework is expected to integrate information from both the input matrix X through the prediction function f and the augmented label matrix through the low-rank regularization to enhance attribute prediction.

The nuclear norm regularization nevertheless is non-smooth. To entail a simple learning procedure, we further exploit a well known identity and encode the low-rank property by introducing two low-dimensional matrices, $\mathbf{U} \in \mathbb{R}^{(L+\hat{L})\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times m}$ ($m < \min(L+\hat{L}, n)$), and replacing $\mathbf{M}$ with $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$. This leads to the following learning formulation:

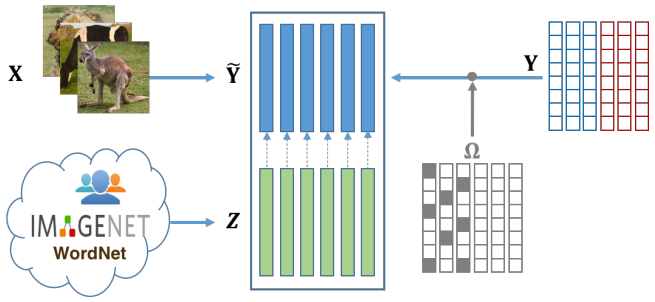

Figure 1: The proposed framework for incomplete attribute learning. It can integrate both observed attribute labels and auxiliary labels for attribute prediction. The red part of $\mathbf{Y}$ denotes the unlabeled part.

$$\min_{\mathbf{W},\mathbf{Y}^u,\mathbf{U},\mathbf{V}} \mathcal{L}(f(\mathbf{X};\mathbf{W}),\mathbf{Y}) + \frac{\alpha}{2}\left\|\begin{bmatrix}\mathbf{Y}^l,\mathbf{Y}^u\\ \mathbf{Z}\end{bmatrix} - \mathbf{U}\mathbf{V}^\top\right\|_F^2 + \frac{\beta}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) + \frac{\gamma}{2}\|\mathbf{W}\|_F^2 \tag{2}$$

3.2 Learning with Incomplete Attribute Labels

As aforementioned, complete attribute annotations are typically difficult to obtain, while incomplete attribute annotations are prevalent. In this case, our label indicator matrix $\mathbf{Y}^\ell$ for the labeled training images is not completely observed. Hence both $\mathbf{Y}^u$ and part of $\mathbf{Y}^\ell$ contain missing labels or unobserved entries. Here we use a mask matrix $\mathbf{\Omega} \in \{0,1\}^{L\times n}$ to indicate the observation status of the corresponding entries of $\mathbf{Y}$. Our proposed transductive learning model above nevertheless can be naturally extended to handle learning with incomplete attribute labels by learning the label matrix for all unobserved entries, which leads to the following formulation:

$$\begin{aligned}\min_{\mathbf{W},\tilde{\mathbf{Y}},\mathbf{U},\mathbf{V}} \quad & \mathcal{L}(f(\mathbf{X};\mathbf{W}),\tilde{\mathbf{Y}}) + \frac{\alpha}{2}\left\|\begin{bmatrix}\tilde{\mathbf{Y}}\\ \mathbf{Z}\end{bmatrix} - \mathbf{U}\mathbf{V}^\top\right\|_F^2 \\ & + \frac{\beta}{2}(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) + \frac{\gamma}{2}\|\mathbf{W}\|_F^2, \\ s.t. \quad & \mathbf{\Omega}\circ\tilde{\mathbf{Y}} = \mathbf{\Omega}\circ\mathbf{Y}; \quad 0 \leqslant \tilde{\mathbf{Y}} \leqslant 1\end{aligned} \tag{3}$$

where $\circ$ denotes the element-wise matrix multiplication. The equality constraints preserve the observed labels in the given label matrix $\mathbf{Y}$. Ideally, our attribute prediction matrix $\tilde{\mathbf{Y}}$ should be an indicator matrix, i.e., $\tilde{\mathbf{Y}} \in \{0,1\}^{L\times n}$. To facilitate convenient optimization, here we relaxed the integer constraints into a continuous range between 0 and 1. The overall learning framework is illustrated in Fig. 1.

In this learning scenario, the low-rank regularization over the augmented label matrix can help to infer the missing attribute labels by exploiting the linear correlations/dependencies between auxiliary labels and attribute labels. For example, by taking the object categories as auxiliary labels, we can infer the attribute "ear" with a high probability if we have already known the object is "cat" with the attribute "head" present. We can not do such reasoning if the object is "bird" because the ear is not visible on the head part of birds under most circumstances.

3.3 Optimization

To obtain a concrete learning problem, we propose to use a linear prediction function f and squared loss function $\mathcal{L}(\cdot,\cdot)$:

$$\mathcal{L}(f(\mathbf{X};\mathbf{W}),\tilde{\mathbf{Y}}) = \left\|\mathbf{W}^\top\mathbf{X} - \tilde{\mathbf{Y}}\right\|_F^2. \tag{4}$$

With this loss function, the learning model in Eqn. (3) is a joint minimization problem over four variables: $\tilde{\mathbf{Y}}, \mathbf{W}, \mathbf{U}$ and $\mathbf{V}$. The objective function is convex in each variable while keeping the other variables fixed. Therefore, we propose to solve this optimization problem using an alternating optimization procedure.

We first initialize $\mathbf{W}$ and $\tilde{\mathbf{Y}}$ by training a linear regression model to predict the partially observed attribute labels:

$$\begin{aligned}(\mathbf{W},\mathbf{W}_z) &= \mathrm{argmin}_{\mathbf{W},\mathbf{W}_z}\left\|\mathbf{\Omega}\circ\left([\mathbf{W};\mathbf{W}_z]^\top\begin{bmatrix}\mathbf{X}\\ \mathbf{Z}\end{bmatrix} - \mathbf{Y}\right)\right\|_F^2, \\ \tilde{\mathbf{Y}} &= (1-\mathbf{\Omega})\circ[\mathbf{W};\mathbf{W}_z]^\top[\mathbf{X};\mathbf{Z}] + \mathbf{\Omega}\circ\mathbf{Y},\end{aligned} \tag{5}$$

where $\mathbf{W}_z$ is the parameter matrix for predicting the auxiliary labels which is only used during the initialization stage. We used the auxiliary labels as inputs in order to achieve a better initialization of $\tilde{\mathbf{Y}}$. We then initialize $\mathbf{U}$ and $\mathbf{V}$ by performing SVD on the augmented label matrix $\left[\tilde{\mathbf{Y}};\mathbf{Z}\right] = \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^\top$, such that

$$\mathbf{U} = \mathbf{P}_{:,1:m}\mathbf{\Sigma}^{\frac{1}{2}}_{1:m,1:m}, \quad \mathbf{V} = \mathbf{Q}_{:,1:m}\mathbf{\Sigma}^{\frac{1}{2}}_{1:m,1:m}. \tag{6}$$

Given these initialization values, we iteratively update the four variables and in each iteration we perform the following two steps. First, given the current value of the parameter matrices $\mathbf{W}, \mathbf{U}$ and $\mathbf{V}$, we optimize $\tilde{\mathbf{Y}}$ in a row-wise manner. The i^{th} row of $\mathbf{Y}$ is updated by solving the following subproblem:

$$\begin{aligned}\tilde{\mathbf{Y}}^*_{i,:} = &\mathrm{argmin}_{\tilde{\mathbf{Y}}_{i,:}}(1+\frac{\alpha}{2})\tilde{\mathbf{Y}}_{i,:}\tilde{\mathbf{Y}}^T_{i,:} - (2\mathbf{W}^\top_{:,i}\mathbf{X} + \alpha\mathbf{U}_{i,:}\mathbf{V}^\top)\tilde{\mathbf{Y}}^\top_{i,:} \\ & s.t. \quad \mathbf{\Omega}_{i,:}\circ\tilde{\mathbf{Y}}_{i,:} = \mathbf{\Omega}_{i,:}\circ\mathbf{Y}_{i,:}; \quad 0 \leqslant \tilde{\mathbf{Y}}_{i,:} \leqslant 1\end{aligned} \tag{7}$$

The above formulation is a constrained quadratic programming problem which can be solve efficiently using a standard quadratic solver. Second, given fixed $\tilde{\mathbf{Y}}$, we use the following closed-form updates for $\mathbf{W}, \mathbf{U}$ and $\mathbf{V}$:

$$\begin{aligned}\mathbf{W} &= (2\mathbf{X}\mathbf{X}^\top + \gamma\mathbf{I})^{-1}2\mathbf{X}\tilde{\mathbf{Y}}^\top, \\ \mathbf{U} &= \begin{bmatrix}\tilde{\mathbf{Y}}\\ \mathbf{Z}\end{bmatrix}\mathbf{V}(\mathbf{V}^\top\mathbf{V} + \frac{\beta}{\alpha}\mathbf{I})^{-1}, \\ \mathbf{V} &= \left[\tilde{\mathbf{Y}}^\top,\mathbf{Z}^\top\right]\mathbf{U}(\mathbf{U}^\top\mathbf{U} + \frac{\beta}{\alpha}\mathbf{I})^{-1}.\end{aligned} \tag{8}$$

This learning process will be stopped if no performance gain is further obtained on the validation set. The overall optimization procedure is summarized in Alg. 1.

Algorithm 1 Optimization procedure

Input: $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{\Omega}$; α, β, γ and m
Initialization:
Initialize $\mathbf{W}$, $\mathbf{Y}$ using Eqn. (5).
Initialize $\mathbf{U}$, $\mathbf{V}$ using Eqn. (6).
repeat
Update $\tilde{\mathbf{Y}} = \tilde{\mathbf{Y}}^*$ by solving Eqn. (7) with fixed $\mathbf{W}$, $\mathbf{U}$ and $\mathbf{V}$
Update $\mathbf{W}$, $\mathbf{U}$ and $\mathbf{V}$ using Eqn. (8) with fixed $\tilde{\mathbf{Y}}$
until no further performance gain

3.4 Mining Auxiliary labels

In this section, we investigate different types of auxiliary labels such as using human annotated category labels and transferring auxiliary labels from external knowledge database.

Human Annotated Object Categories

In many tasks, object category annotations are usually available in the training data. Moreover, from the presence probability of the pre-defined attributes calculated with images belonging to the same category, as shown in Fig. 2, we found that images from similar object categories usually share more common attributes. Therefore, we first investigate using object categories as auxiliary labels to infer the missing part of attributes. In this case, $\mathbf{Z}$ is a sparse matrix with only one non-zero element for each column.

For conventional attribute learning problem, the object category annotations for unseen data are usually not observed. Therefore, we choose to train a category prediction model first by using the category supervision of the seen data. Then the learned model is used to produce the auxiliary category labels for the unseen data as $\mathbf{Z}^u$. Here a ridge regression model is trained using the data $\mathbf{X}^l$ with known object category annotations $\mathbf{Z}^\ell$ and the annotations for $\mathbf{X}^u$ can be then produced as following:

$$\mathbf{Z}^u = \mathbf{Z}^l {\mathbf{X}^l}^\top (\mathbf{X}^l {\mathbf{X}^l}^\top + \lambda \mathbf{I})^{-1} \mathbf{X}^u, \tag{9}$$

where λ is the hyper parameter for the L_2 model parameter regularization in a linear regression model. The auxiliary matrix in Eqn. (3) can be formed as $\mathbf{Z} = \left[\mathbf{Z}^l, \mathbf{Z}^u\right]$.

Knowledge Transferring from External Database

In addition to using human annotated category labels, we also investigate auxiliary labels from the external dataset. In particular, we propose to use Large Scale Visual Recognition Challenge 2012 (ILSVRC 2012) [Russakovsky *et al.*, 2015] as the external database. It contains 1.2 million images and 1000 object categories. We consider the ILSVRC 2012 dataset as the source domain and the dataset for attribute learning as the target domain. Only part of the categories defined in the target domain may appear in the source domain. By using a base model (e.g. AlexNet [Krizhevsky *et al.*, 2012]) pre-trained on ILSVRC 2012, we can extract the category prediction $\mathbf{S} \in \mathbb{R}^{c \times n}$ for all the images $\mathbf{X}$ in the target domain, where $c = 1000$ denotes the number of source domain object categories and the sum of each column of $\mathbf{S}$ equals to one. Since the source domain contains much more object categories than the target domain, it is not efficient to directly use all the posterior probabilities as auxiliary labels.

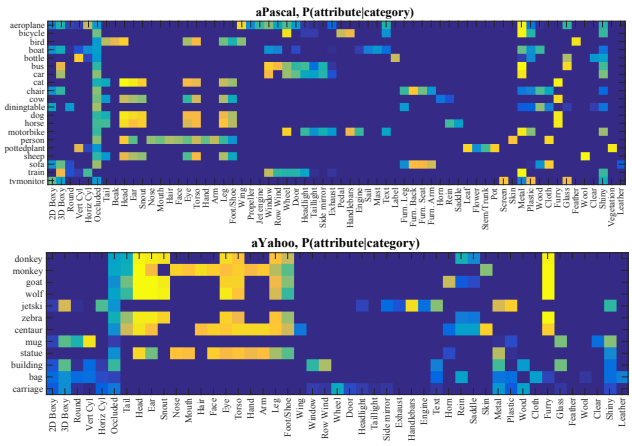

Figure 2: The presence probability of attributes for different object categories on aPascal and aYahoo datasets (Yellow means high probability).

Therefore, we propose the following two ways to make full use of the external knowledge.

Semantic Pooling. We first propose Semantic Pooling to select the *most relevant* object categories from the source domain as the auxiliary labels for attribute learning. By summing up the presence probabilities of each source domain object category on the target domain images, such as $\tilde{\mathbf{S}}_j = \sum_i \mathbf{S}_{j,i}$, we can find the sum probabilities demonstrate a long-tail distribution which means only a small proportion of object categories from the source domain are relevant to the target dataset. Therefore, we propose to pool the most relevant object categories as the auxiliary labels. We use $\mathbf{R} = \{j | \tilde{\mathbf{S}}_j > t\}$ to denote the indices of the selected object categories, where t is the threshold for category selection. Then we can produce the auxiliary category label matrix as $\mathbf{Z} = \mathbf{S}(\mathbf{R}, :)$ and further normalize $\mathbf{Z}$ to keep each column sum up to one.

Semantic Propagation. Instead of only using part of the object categories of the source domain, we also try to directly propagate the posterior probabilities of all the source object categories to the target object categories. We model the similarity between the object categories of the source domain and target domain using WordNet hierarchy [Fellbaum, 1998]. We denote the i^{th} object in the source domain and the j^{th} object in the target domain as $\mathbf{O}_i^s$ and $\mathbf{O}_j^t$ respectively. The similarity of the two categories is measured by the Wu-Palmer Similarity [Bird *et al.*, 2009]. It is based on the depth of the two senses in the taxonomy and their Least Common Subsumer, and is calculated as $\mathcal{K}(\mathbf{O}_i^s, \mathbf{O}_j^t) = 2 * Depth(LCS(\mathbf{O}_i^s, \mathbf{O}_i^t)) / (Depth(\mathbf{O}_i^s) + Depth(\mathbf{O}_j^t))$. The propagation matrix can then be constructed as following:

$$\mathbf{T}_{i,j} = \frac{exp(\rho \mathcal{K}(\mathbf{O}_i^s, \mathbf{O}_j^t)^2)}{\sum_{k=1}^{\hat{L}} exp(\rho \mathcal{K}(\mathbf{O}_i^s, \mathbf{O}_k^t)^2)}, \tag{10}$$

where ρ is a parameter to be specified. The auxiliary label matrix can be obtained as $\mathbf{Z} = \mathbf{T}^\top \mathbf{S}$ by propagating the posterior probabilities from source domain to target domain.

Table 1: Detailed information for the datasets.

Dataset	# images	# attributes	# objects
aPascal	12785	64	20
aYahoo	2644	47	12
Imagenet attribute	9600	25	384

4 Experiments

4.1 Experimental Setting

Datasets. We conducted experiments on three real-world datasets for attribute learning. aPascal [Farhadi *et al.*, 2009] contains 6430 training images and 6355 testing images from Pascal VOC 2008 challenge. Each image comes from twenty object categories and is annotated with 64 binary attribute labels. aYahoo [Farhadi *et al.*, 2009] contains 2644 images belonging to twelve object categories. Each image is annotated with the same 64 binary attributes as the *aPascal* dataset. By discarding the attributes with no positive data, we finally get 47 attributes to conduct experiments. INA (ImageNet Attributes [Russakovsky and Fei-Fei, 2010]) contains 9,600 images across 384 categories. Each image is annotated with 25 binary attributes. The information about the three datasets are summarized in Table 1.

Experimental Setup. For the aPascal dataset, we use the default {train, test} split and separate half the training data for validation. For aYahoo and INA, we randomly split the dataset into three subsets with equal size for training, validating and testing. We used the Convolutional Neural Networks (CNN) [Donahue *et al.*, 2014] to extract 4096 DeCAF features for each image within the provided bounding box area. The performance of attribute predictors are measured by *mAUC* (mean Area Under ROC) and *mAP* (mean Average Precision) to reflect the average performance of all the attributes. To simulate the incomplete attribute learning setting, we randomly used $\{10, 20, 30, 40, 50\}$ percent of the annotated labels for model training. We compared all methods using the same data setting, randomly sampled the observed attribute labels and repeated each experiment five times.

Comparison Methods. In the experiments, we compare the proposed approach with the following methods: (1) the mixed graph method for multi-label learning with missing labels (*ML-MG*) [Wu *et al.*, 2015]; (2) the unified multiplicative framework for attribute Learning (*UMF*) [Liang *et al.*, 2015]; (3) the concatenation methods with multiple input information (*Concat*); and (4) the baseline binary relevance method (*BR*). The first two methods are state-of-the-art methods for multi-label learning with incomplete labels and visual attribute learning respectively. *ML-MG* incorporates instance level similarity and label dependencies to handle missing labels. *UMF* integrates object recognition into conventional attribute learning model in a multiplicative way. The *Concat* method also exploits object labels; it concatenates the image features and the auxiliary object labels together as the input data. Comparing with *UMF*, *Concat* leverages multiple information in an additive way. *BR* is a widely used method for multi-label classification. We independently train logistic regression model for each binary attribute for *BR*.

We used the open-source code of *ML-MG* and *UMF* to conduct the experiments. For all the methods, we conducted parameter selection based on the performance on the validation set. For our proposed approach, we select the trade-off parameters α from $\{10^{-2}; 10^{-1}; 1; 10; 100\}$, select β from $\{10^{-5}; 10^{-4}; 10^{-3}; 10^{-2}\}$ while setting β and γ to be equal.

4.2 Experiment Results

Incomplete Attribute Learning

In this section, we take human annotated object categories as the auxiliary labels. We use the ground-truth category labels for seen data and acquire object category prediction for unseen data based on Eqn. (9). From the results in Fig. 3, we can see that the approaches integrated with auxiliary labels (*UMF*, *Concat* and the proposed method) perform better than the other methods especially when the observed attribute labels are rare. This demonstrates the effectiveness of using auxiliary labels on attribute prediction. By comparing different ways for leveraging auxiliary labels, we find that *Concat* works not very effectively since it trends to fall behind *BR* on aPascal and aYahoo with increasing number of observed attribute labels. The main difference between *Concat* and our proposed method is *Concat* uses the auxiliary labels to directly infer the missing part of attributes without considering the observed part of attributes. Compared with *Concat* and the proposed method, *UMF* works well on aPascal and aYahoo but fails on INA which has many more categories. *ML-MG* seems to not perform well on attribute learning tasks. The main reason can be that attributes do not have the semantic label hierarchy such as "animal→horse" and "plant→grass" which are commonly presented in multi-label learning problem. Moreover, the co-occurrence of attributes is hard to exploit if the relationship with attributes and objects are not well considered.

Our proposed approach constantly outperforms the other comparison methods based on the mAP evaluation metric on all the three datasets especially when only a small portion of attribute labels are observed. By observing 10 percent of attributes on the *aPascal* dataset, our proposed method improves the state-of-the-art performance about 2% according to both *mAP* and *mAUC*. For the aYahoo dataset, we achieve the best *mAP* performance but falls behind *UMF* based on *mAUC*. Since the presences of most attributes are much less than their absences, attribute learning usually suffers from data imbalance problem. Referring to [Davis and Goadrich, 2006], using an evaluation metric of Precision-Recall curve is more reasonable than the ROC curve to measure the comparison methods on the imbalance learning task.

Mining Auxiliary Labels

We conduct experiments by taking advantage of the external database. The source domain is specified to be the 1000 object categories defined in ILSVRC 2012 dataset. We extract the posterior probabilities of source domain object categories on the input images using two base networks: AlexNet [Donahue *et al.*, 2014] and VGG-16 [Simonyan and Zisserman, 2014]. Then we use the two methods proposed in Sec. 3.4 to conduct the auxiliary label matrix $\mathbf{Z}$. For semantic propagation, we manually map the object categories from both ILSVRC 2012 and the three benchmark datasets into the WordNet hierarchy. Then we can measure the similarity be-

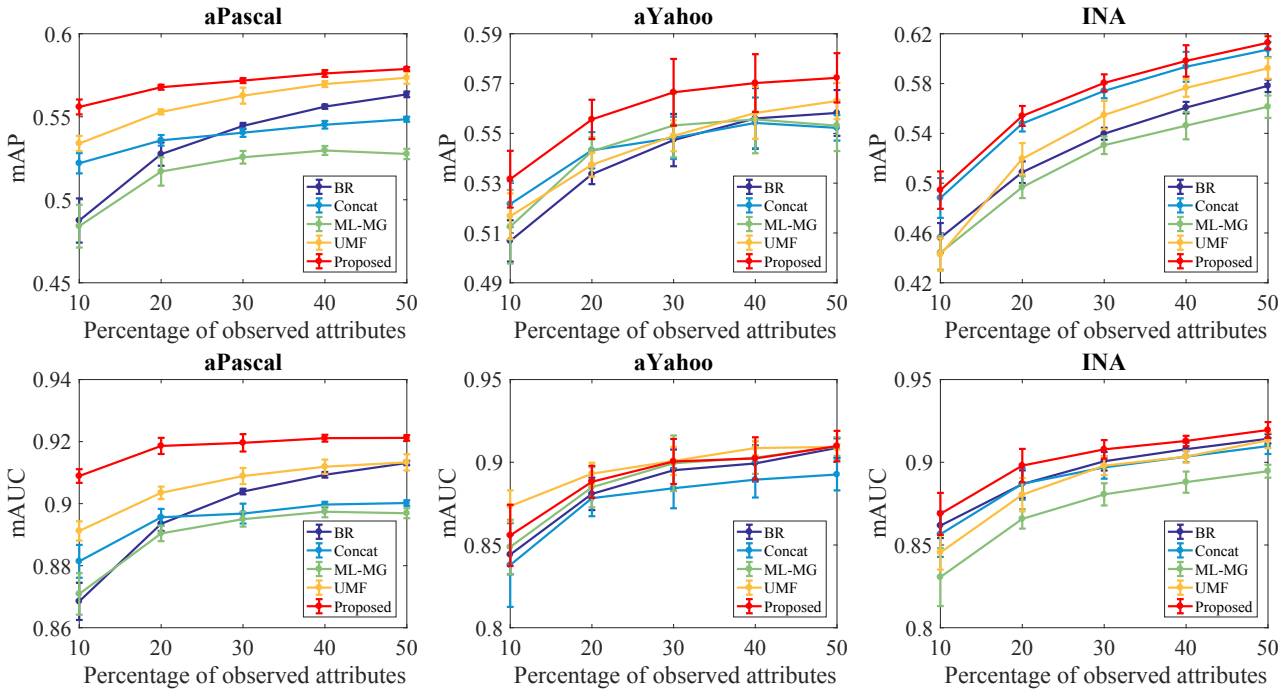

Figure 3: The performance of different comparison methods on the three benchmark datasets with incomplete attribute labels.

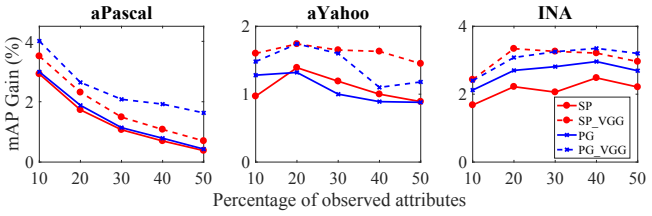

Figure 4: Knowledge Transferring from External Database.

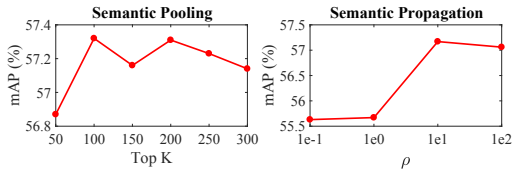

Figure 5: Hyperparameters tuning on aPascal Validation.

tween two arbitrary object categories and compute the propagation matrix based on Eqn. (10).

We calculated the performance gain of using auxiliary labels over the binary relevance method (*BR*). As shown in Fig. 4, mining auxiliary labels is always helpful for learning from incomplete attributes. By comparing two methods of generating auxiliary labels, Semantic Propagation (PG) achieves better performance than Semantic Pooling (SP) on aPascal and INA. This shows the effectiveness of leveraging WordNet taxonomy. Comparing with using AlexNet as base network, using VGG alway shows better performance. This is reasonable as VGG achieves a lower error rate than AlexNet on the ILSVRC 2012 dataset. We conducted the experiments by setting different values of hyperparameters with observing 10% attributes on aPascal. For Semantic Pooling, we modify the threshold to pool the top K object categories of source domain. As shown in Fig.5, the performance of Semantic Pooling starts to drop when more object categories are involved as auxiliary labels. For Semantic Propagation, choosing a larger value for ρ can achieve better performance.

Comparing Fig. 3 and Fig. 4, the proposed method achieves better performance by using human annotated object categories rather than mining auxiliary labels. The main reason is part of the objects from target datasets are missing on ILSVRC 2012 dataset though the latter contains many more object categories. However, mining auxiliary labels is still promising as it does not need any human annotations which dramatically decreases the cost of labeling.

5 Conclusion

We proposed a novel transductive learning method by integrating auxiliary labels for incomplete attribute learning. By modeling the relationship of attributes and auxiliary labels, the missing attributes can be recovered effectively. The proposed model can be solved efficiently by alternatively optimizing constrained quadratic programming problems and parameter updating in closed form solutions. In addition, we investigate different ways to acquire auxiliary labels. By taking the auxiliary labels as the human annotated object category labels, our proposed method can achieve the state-of-the-art performance on three widely used datasets. Moreover, the auxiliary labels transferred from a large scale dataset can also improve the performance without adding extra human cost.

Acknowledgments

Research supported by China Scholarship Council (No. 201604910935), Natural Science Foundation of China (No. 61390515) and the Canada Research Chairs program.

References

[Bird *et al.*, 2009] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc", 2009.

[Chen *et al.*, 2013] Minmin Chen, Alice X Zheng, and Kilian Q Weinberger. Fast image tagging. In *Proc. of ICML*, pages 1274–1282, 2013.

[Chen *et al.*, 2014] Lin Chen, Qiang Zhang, and Baoxin Li. Predicting multiple attributes via relative multi-task learning. In *Proc. of CVPR*, pages 1027–1034. IEEE, 2014.

[Davis and Goadrich, 2006] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proc. of ICML*, pages 233–240. ACM, 2006.

[Donahue *et al.*, 2014] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proc. of ICML*, pages 647–655, 2014.

[Fang *et al.*, 2015] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, et al. From captions to visual concepts and back. In *Proc. of CVPR*, pages 1473–1482, 2015.

[Farhadi *et al.*, 2009] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Proc. of CVPR*, pages 1778–1785. IEEE, 2009.

[Fellbaum, 1998] Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

[Frome *et al.*, 2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Proc. of NIPS*, pages 2121–2129, 2013.

[Huang *et al.*, 2015] Sheng Huang, Mohamed Elhoseiny, Ahmed Elgammal, and Dan Yang. Learning hypergraph-regularized attribute predictors. In *Proc. of CVPR*, pages 409–417, 2015.

[Hwang and Sigal, 2014] Sung Ju Hwang and Leonid Sigal. A unified semantic embedding: Relating taxonomies and attributes. In *Proc. of NIPS*, pages 271–279, 2014.

[Jayaraman and Grauman, 2014] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. In *Proc. of NIPS*, pages 3464–3472, 2014.

[Jayaraman *et al.*, 2014] Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *Proc. of CVPR*, pages 1629–1636. IEEE, 2014.

[Kovashka *et al.*, 2012] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *Proc. of CVPR*, pages 2973–2980. IEEE, 2012.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. of NIPS*, pages 1097–1105, 2012.

[Lampert *et al.*, 2014] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 36(3):453–465, 2014.

[Liang *et al.*, 2015] Kongming Liang, Hong Chang, Shiguang Shan, and Xilin Chen. A unified multiplicative framework for attribute learning. In *Proc. of ICCV*, pages 2506–2514, 2015.

[Liang *et al.*, 2016] Kongming Liang, Hong Chang, Shiguang Shan, and Xilin Chen. Attribute conjunction learning with recurrent neural network. In *Proc. of ECML-PKDD*, pages 345–360. Springer, 2016.

[Lu *et al.*, 2016] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proc. of ECCV*, pages 852–869. Springer, 2016.

[Lu, 2016] Yao Lu. Unsupervised learning on neural network outputs: with application in zero-shot learning. In *Proc. of IJCAI*, 2016.

[Russakovsky and Fei-Fei, 2010] Olga Russakovsky and Li Fei-Fei. Attribute learning in large-scale datasets. In *Proc. of ECCV Workshop*, 2010.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015.

[Simonyan and Zisserman, 2014] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[Sun *et al.*, 2010] Yu-yin Sun, Yin Zhang, and Zhi-hua Zhou. Multi-label learning with weak label. In *Proc. of AAAI*. Citeseer, 2010.

[Wang and Ji, 2013] Xiaoyang Wang and Qiang Ji. A unified probabilistic approach modeling relationships between attributes and objects. In *Proc. of ICCV*, 2013.

[Wu *et al.*, 2013] Lei Wu, Rong Jin, and Anil K Jain. Tag completion for image retrieval. *IEEE TPAMI*, 35(3):716–727, 2013.

[Wu *et al.*, 2015] Baoyuan Wu, Siwei Lyu, and Bernard Ghanem. Ml-mg: multi-label learning with missing labels using a mixed graph. In *Proc. of ICCV*, 2015.

[You *et al.*, 2016] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proc. of CVPR*, pages 4651–4659, 2016.

[Yu *et al.*, 2014] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit S Dhillon. Large-scale multi-label learning with missing labels. In *Proc. of ICML*, pages 593–601, 2014.

[Zhao and Guo, 2015] Feipeng Zhao and Yuhong Guo. Semi-supervised multi-label learning with incomplete labels. In *Proc. of IJCAI*, pages 4062–4068. AAAI Press, 2015.