A Unified Multiplicative Framework for Attribute Learning

Kongming Liang[1,2], Hong Chang[1], Shiguang Shan[1], Xilin Chen[1]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
{kongming.liang, hong.chang, shiguang.shan, xilin.chen}@vipl.ict.ac.cn

Abstract

Attributes are mid-level semantic properties of objects. Recent research has shown that visual attributes can benefit many traditional learning problems in computer vision community. However, attribute learning is still a challenging problem as the attributes may not always be predictable directly from input images and the variation of visual attributes is sometimes large across categories. In this paper, we propose a unified multiplicative framework for attribute learning, which tackles the key problems. Specifically, images and category information are jointly projected into a shared feature space, where the latent factors are disentangled and multiplied for attribute prediction. The resulting attribute classifier is category-specific instead of being shared by all categories. Moreover, our method can leverage auxiliary data to enhance the predictive ability of attribute classifiers, reducing the effort of instance-level attribute annotation to some extent. Experimental results show that our method achieves superior performance on both instance-level and category-level attribute prediction. For zero-shot learning based on attributes, our method significantly improves the state-of-the-art performance on AwA dataset and achieves comparable performance on CUB dataset.

1. Introduction

Attributes are namable properties of objects which are observable from visual images. They can be annotated in either instance-level or category-level (e.g. images belongs to the same category share common attribute annotation). Beyond traditional object recognition, learning attributes can provide fine-grained descriptions, such as the holistic perception (e.g., color, shape, etc.) and presence or absence of local parts for images. As a mid-level semantic cue, they can bridge the gap between low-level features and high-level categorization. Recent research has verified that attributes can benefit many traditional learning problems (e.g., image search [15], object recognition [22] and face verification [16]). Moreover, they provide a proper way to address *zero-shot classification* [17] by transferring from seen classes to unseen classes.

Direct attribute prediction methods [8, 17] train a group of binary classifiers from image-attribute pairs, one individually for each attribute. Fig. 1 (a) illustrates the direct attribute learning method. During test stage, the learned classifiers are applied to predict which subset of attributes the input image may have. Though these methods achieve a relatively good performance in predicting attribute and recognizing unseen categories, there are some obvious limitations as following:

1. Correlation between attributes are ignored. Naturally, attribute as properties of objects are correlated with each other, therefore it is more appropriate to learn all the attributes jointly, such as sharing attribute-specific parameters or common semantic representations.

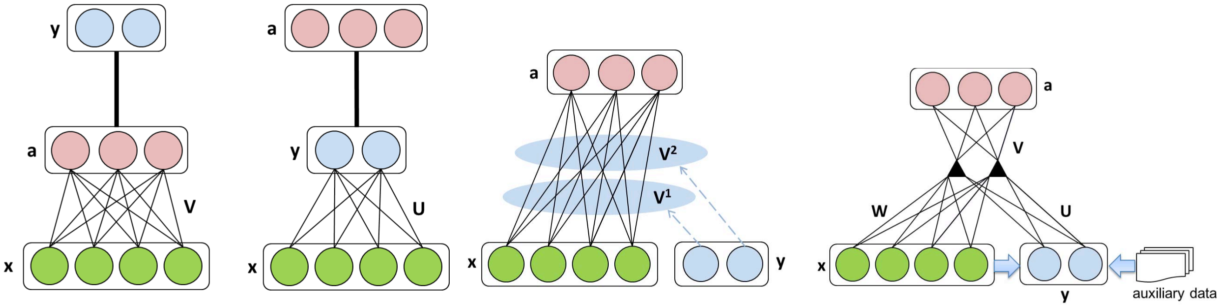
2. Some attributes are hard or even unable to predict based on visual appearances. For example, it is impossible to infer color-relevant attribute from an gray image input or predict whether an animal is fast or slow based on an still image.

3. Negative attribute correlation between object and scene. For weakly supervised attribute learning[17], the input image contains both object and scene. It happens sometimes that the scene has some attributes that are negatively related to object attributes. For example, traditional attribute classifier may predict a polar bear swimming in the ocean to have blue attribute.

4. Different visual attribute manifestations vary across categories. For example, the same attribute concept “fluffy” varies considerably between dog and towel[6].

Some methods have been proposed to tackle the above problems to some extent. Lampert et al. [17] propose a method to indirectly predict attributes by transferring

1550-5499/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCV.2015.288



(a) (b) (c) (d)

Figure 1. Models for attribute learning. (a) direct attribute prediction model; (b) indirect attribute prediction model; (c) category-sensitive attribute learning model; (d) our proposed multiplicative model. $\mathbf{x}$, $\mathbf{a}$ and $\mathbf{y}$ denote image, attribute and label vectors respectively. $\mathbf{W}$, $\mathbf{U}$, $\mathbf{V}$ are model parameters. Bold lines mean known relationship.

knowledge between classes which can infer some attributes which are unable to detect directly. Fig. 1 (b) illustrates the indirect attribute learning method. Jayaraman et al. [12] and Chen et al. [6] formulate attribute learning in regularization-based multi-task learning framework, where each subtask corresponds to learning one attribute. Wang et al. [21] use Bayesian network to enhance attribute prediction by leveraging the statistical relationships between attributes and objects. Huang et al. [10] model attribute learning as a supervised hypergraph cut problem to learn attributes jointly and exploiting class information.

In this paper, we propose a unified multiplicative multi-task learning framework to address all the above problems. Fig.1 (d) illustrates our model, where the image and category vectors in the unified common space interact multiplicatively to predict the attributes. During the training stage, all parameters are learned to automatically balance the information to be leveraged. In sum, the main advantages of our proposed method are as follows: (1) By projecting input images and categories into a latent common space, factors correlated to all attributes are disentangled and multiplied for attribute prediction. (2) Our method can leverage category information to infer attribute when the latter is hard or unable to be predicted. In addition, when negative correlation exists, the scene used as context information is helpful to predict category. In this way, scene information can be converted into positive cues for indirect attribute prediction. (3) The attribute classifier in our method are instance-specific and can be decomposed into a linear combination of category-specific attribute classifiers. Thus, it gives a finer attribute description than conventional attribute classifiers and can be transferred to unseen class more easily. (4) Experimental results show that our method achieves superior performance in not only attribute prediction but also zero-shot learning.

The rest of this paper is organized as follows. We first introduce some related works in the following section. In Section 3, we present the unified multiplicative model in detail. Experimental results are then shown in Section 4.1. Finally, Section 5 concludes the paper.

2. Related Work

Semantic embedding. Our method can be viewed as a unified semantic embedding for images, attributes, and categories. Akata et al. [1, 2] use category-level attribute vectors as class embedding and model the relationship between images and class embedding by a bilinear function. Each column of the function parameters can be interpreted as an attribute classifier which is shared by all the categories. Since visual attribute manifestations vary across categories, the assumption of parameter sharing is not appropriate, especially when the inter-class variation is large. Hwang et al. [11] explicitly embed all semantic entities including attributes and supercategories into the same space. Then an object can be represented as linear combination of the semantic entities. Without considering the variation of visual attribute manifestations, the attribute embedding is also shared across categories. In addition, they mainly learn the unified semantic space for better multi-class classification accuracy, while we focus on maximum likelihood estimation of logistic regression model for attribute prediction. Fu et al. [9] propose a framework called transductive multi-view embedding to tackle the projection domain shift problem in zero-shot learning.

Multi-task learning. Our method is in the multi-task learning framework, where each task corresponds to learning one semantic attribute. Jayraman et al. [12] propose to decorrelate visual attributes by group lasso regularization based on attribute group information. In this way, feature sharing is encouraged in the same attribute group and feature competition is promoted across different groups. Chen et al. [6] adopt a robust regularization scheme to detect outlier attribute and leverage correlations among attributes.

Multiplicative models. Multiplicative models are effective in relating separate underlying factors in data. [20]

proposes a general framework of multiplicative multi-task learning which decomposes the model parameters of each task into a multiplication of two components, the cross-task component and the task-specific one. Our method is different from this work as we use a multiplication of three components to model the third-order relationship between image, category and attribute. And the cross-task component in our method is not under the assumption of being a vector. [13] presents a multimodal neural language model in a multiplicative form, where images are used for gating word representations. In our method, the category-level information can be considered as the gate for attribute prediction. Our formulation on the relationship between gated inputs and attributes is different from that of [13], which makes use a known language model. [14] also uses multiplicative models to learn the third-order relationship. However, in [14], attributes are provided to learn the conditional word similarity, while our model predicts category-sensitive attributes by leveraging category information from a classification model.

Usage of category information. Lampert et al. [17] propose a method to indirectly predict attributes by transferring knowledge between classes. However, this method can not predict instance-level attributes and totally ignores the low-level visual cue. Wang et al. [21] propose a unified probabilistic model to capture the class-dependent and class-independent attribute relationships, which benefit both attribute prediction and object recognition. [5] models high-order relationship between attribute and category to predict category-sensitive attributes and infer unseen category-attribute pairs by using tensor completion based on a sparse set of category-specific attribute classifiers. Fig. 1 (c) illustrates the method. Huang et al. [10] propose to model the attribute learning as a supervised hypergraph cut problem and consider it as a multi-graph cut problem to incorporate category information.

3. Our Proposed Method

We begin by introducing some notations. Assume there are T attributes to be predicted, each of them being considered as one task in the multi-task learning framework. Suppose there are N labeled training images, $\{\mathbf{x}_i, \mathbf{a}_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^D$ denotes the D-dimensional image feature vector, and $\mathbf{a}_i \in \{0,1\}^T$ indicates the absence or presence of all binary attributes. Each image $\mathbf{x}_i$ has a class label vector $\mathbf{y}_i \in \mathbb{R}^C$, where C is the number of classes. The training images can be expressed in matrix form as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D\times N}$, similarly for the attribute matrix $\mathbf{A} \in \mathbb{R}^{T\times N}$ and class label matrix $\mathbf{Y} \in \mathbb{R}^{C\times N}$.

3.1. Multiplicative Attribute Learning Model

We transform training images and their class labels into a shared feature space, where the latent factors correlated to attributes are disentangled. Suppose linear mappings for images $\mathbf{X}$ and labels $\mathbf{Y}$ from their original spaces to the latent feature space, which are parameterized respectively by $\mathbf{W} \in \mathbb{R}^{F\times D}$ and $\mathbf{U} \in \mathbb{R}^{F\times C}$. F is the dimensionality of the latent feature space. Then, $\mathbf{W}\mathbf{x}_i$ and $\mathbf{U}\mathbf{y}_i$ represent the feature representations of image $\mathbf{x}_i$ and its class information in the latent space.

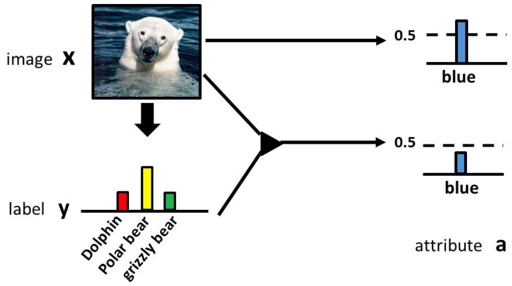


Figure 2. Illustration of our method for predicting attribute "blue". Category information helps to address the negative correlation problem. 0.5 denotes the decision boundary for attribute prediction.

In multi-task learning framework, the t^{th} $(t = 1, \ldots, T)$ task corresponds to learning a binary classifier for the t^{th} attribute. Let $\mathbf{v}_t \in \mathbb{R}^F$ denotes the parameters of the t^{th} classifier in the latent space. Different from traditional attribute learning methods, we relate all parameters using a multiplication model for attribute classification. Formally, the discriminant function of the t^{th} attribute of the object in image $\mathbf{x}_i$ is defined as follows:

$$\begin{aligned} f(\mathbf{x}_i, \mathbf{y}_i, t) &= (\mathbf{v}_t)^T((\mathbf{U}\mathbf{y}_i) \odot (\mathbf{W}\mathbf{x}_i)) \quad (1) \\ &= \langle \mathbf{v}_t, \mathbf{U}\mathbf{y}_i, \mathbf{W}\mathbf{x}_i \rangle, \quad (2) \end{aligned}$$

where the operator $\odot$ denotes element-wise multiplication, i.e., $((\mathbf{U}\mathbf{y}_i) \odot (\mathbf{W}\mathbf{x}_i))_k = (\mathbf{U}\mathbf{y}_i)_k(\mathbf{W}\mathbf{x}_i)_k$, $k = 1, \ldots, F$. In the above equation, the discriminant function is a multiplication (inner product) of three components. The component $\mathbf{W}\mathbf{x}_i$ means to learn a better visual representation for image $\mathbf{x}_i$ to facilitate attribute classification. The component $\mathbf{U}\mathbf{y}_i$ is used as a gate for the attribute classifier $\mathbf{v}_t$ to transfer knowledge from category information. Actually, the category-level information is an important factor for attribute learning, as the visual appearances of attributes are usually class-sensitive. Besides, for some attributes which are hard to predict based on visual cues, we can infer them from category information. Moreover, category-level information may be helpful to address the negative correlation problem, as illustrated in Figure 2. The model parameters $\Phi = \{\mathbf{W}, \mathbf{U}, \mathbf{V}\}$ are shared across all images and tasks. During training stage, all the parameters will be learned to automatically decide how to leverage image, attribute and category information.

Based on the discriminant function defined above, we can make use of logistic regression model to jointly learn all attributes. The loss function is expressed as the negative log likelihood:

$$L(\mathbf{X},\mathbf{Y},\mathbf{A};\Phi) = \frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}[-a_{ti}log(g(f(\mathbf{x}_i,\mathbf{y}_i,t))) \\ -(1-a_{ti})log(1-g(f(\mathbf{x}_i,\mathbf{y}_i,t)))], \tag{3}$$

where Φ is the set of parameters to be learned. a_{ti} indicates the presence or absence of the t^{th} attribute for image $\mathbf{x}_i$. $g(\mathbf{x})$ is a sigmoid function.

The final objective of our multiplicative model takes the following form:

$$J = L(\mathbf{X},\mathbf{Y},\mathbf{A};\Phi)+\lambda_1\Omega(\mathbf{W})+\lambda_2\Omega(\mathbf{U})+\lambda_3\Omega(\mathbf{V}), \tag{4}$$

where $\Omega(\cdot)$ is a regularizer on the mapping matrices and attribute classifier. The parameters λ_1, λ_2 and λ_3 are used to trade the relative influence of the three regularization terms. In this paper, we choose the squared Frobenius norm as the form of $\Omega(\cdot)$.

3.1.1 Category-specific attribute classifier

The discriminant function for the t^{th} attribute of $\mathbf{x}_i$, as expressed in Eqn. (1), can also be written as:

$$f(\mathbf{x}_i,\mathbf{y}_i,t) = ((\mathbf{U}\mathbf{y}_i)\odot\mathbf{v}_t)^T(\mathbf{W}\mathbf{x}_i) \tag{5}$$

$$= \left(\sum_{j=1}^{C} y_{ji}(\mathbf{u}_j\odot\mathbf{v}_t)\right)^T(\mathbf{W}\mathbf{x}_i). \tag{6}$$

Here $\mathbf{u}_j$ is the j^{th} column of $\mathbf{U}$, and y_{ji} is the binary category label which indicates whether x_i belongs to object category j. $(\mathbf{u}_j\odot\mathbf{v}_t)$ acts as the t^{th} attribute classifier which is specific for the object category j. In our method, each input image $\mathbf{x}_i$ is transformed (by $\mathbf{W}$) into the latent feature space, and its attributes are predicted by the category-specific attribute classifiers.

With the training samples, we train a softmax multi-class object classifier by minimizing the following loss function:

$$L_Y(\mathbf{X},\mathbf{Y};\Theta) \\ = \frac{1}{N}\sum_{j=1}^{C}\sum_{i=1}^{N}-1\{y_{ji}=1\}\frac{exp(\boldsymbol{\theta}_j^T\mathbf{x}_i)}{\sum_{k=1}^{C}exp(\boldsymbol{\theta}_k^T\mathbf{x}_i)}, \tag{7}$$

where θ_j is the classifier parameter for object category j. We use the given training set to learn the parameters with a weight decay regularization. At test stage, the category probabilities of a test image $\mathbf{x}$ can be estimated as follows:

$$\tilde{\mathbf{y}} = \begin{bmatrix}\tilde{y}_1\\ \tilde{y}_2\\ \vdots\\ \tilde{y}_C\end{bmatrix} = \frac{1}{\sum_{j=1}^{C}exp(\boldsymbol{\theta}_j^T\mathbf{x})}\begin{bmatrix}exp(\boldsymbol{\theta}_1^T\mathbf{x})\\ exp(\boldsymbol{\theta}_2^T\mathbf{x})\\ \vdots\\ exp(\boldsymbol{\theta}_C^T\mathbf{x})\end{bmatrix}. \tag{8}$$

With the estimated category information, we can predict the attributes of $\mathbf{x}$ by marginalizing the category labels as follows:

$$p(a_t=1|\mathbf{x};\Phi) = \sum_{j=1}^{C}\tilde{y}_j g(f(\mathbf{x},\mathbf{e}_j,t)), \tag{9}$$

where $\mathbf{e}_j$ denotes a vector with only one nonzero coordinate of value 1 in j^{th} position.

3.1.2 Instance-specific attribute classifier

Besides utilizing a separately trained model, we can also jointly train the multi-class classification model (Eqn. 7) and attribute classifiers. In that case, the objective of our multiplicative framework can be re-expressed as:

$$\tilde{J} = L(\mathbf{X},\tilde{\mathbf{Y}},\mathbf{A};\Phi)+\beta L_Y(\mathbf{X},\mathbf{Y};\Theta)+\lambda_1\Omega(\mathbf{W}) \\ +\lambda_2\Omega(\mathbf{U})+\lambda_3\Omega(\mathbf{V})+\lambda_4\Omega(\Theta). \tag{10}$$

Note that the softmax outputs of object categories replace the binary category labels in the first loss term. After joint training, we obtain instance-specific attribute classifiers for $\mathbf{x}_i$:

$$(\mathbf{U}\tilde{\mathbf{y}}_i)\odot\mathbf{v}_t = \sum_{j=1}^{C}\tilde{y}_{ji}(\mathbf{u}_j\odot\mathbf{v}_t). \tag{11}$$

From the equation above, we can see that the classifier for the t^{th} attribute is dependent on not only the category but also the input image itself. Actually, it is a linear combination of all the category-specific attribute classifiers. We can empirically prove that instance-specific classifiers are superior to category-specific classifiers in attribute prediction.

For zero-shot learning, the instance-specific attribute classifier for an image from an unseen category can be estimated by the category-specific attribute classifiers of all the seen categories.

3.2. Optimization

Traditional multiplicative models are optimized by alternating optimization algorithms. It converts the original problem into several subproblems with respect to each parameter and optimizes one parameter in a subproblem with others being fixed. Such optimization process is alternated until the model converges to a local minimum, as analyzed in [3]. In our work, we also use alternate optimization to minimize the objective function in Eqn. (4). The overall algorithm is described in Algorithm 1.

As presented in the algorithm, instead of randomly initializing $\mathbf{W}$ and $\mathbf{V}$, we initialize them with the SVD decomposition of traditional logistic regression classifier parameters. The derivative of the objective function with respect to

Algorithm 1 Alternating optimization for UMF

```
Input: image feature $\mathbf{X}$, category information $\mathbf{Y}$, attribute labels $\mathbf{A}$, latent space dimension $F$, and balance parameters $\lambda_1$, $\lambda_2$, $\lambda_3$
Output: $\mathbf{U}, \mathbf{V}, \mathbf{W}$
Train logistic regression classifiers for attribute learning and
    get the parameter matrix $\mathbf{E}$
Do SVD decomposition for $\mathbf{E} = \mathbf{PSQ}^T$
Initialize $\mathbf{W}_0 = \mathbf{S}^{\frac{1}{2}}_{1:F,1:F}\mathbf{P}^T_{:,1:F}$, $\mathbf{V}_0 = \mathbf{S}^{\frac{1}{2}}_{1:F,1:F}\mathbf{Q}^T_{:,1:F}$
    and $\mathbf{U}_0$ with random value
Set $t = 0$
repeat
    L-BFGS optimization for $\mathbf{U}^*$ with fixed $\mathbf{W}_t$,$\mathbf{V}_t$
    Update $\mathbf{U}^{t+1} = \mathbf{U}^*$
    L-BFGS optimization for $\mathbf{W}^*$ with fixed $\mathbf{V}_t$,$\mathbf{U}_{t+1}$
    Update $\mathbf{W}^{t+1} = \mathbf{W}^*$
    L-BFGS optimization for $\mathbf{V}^*$ with fixed $\mathbf{U}_{t+1}$,$\mathbf{W}_{t+1}$
    Update $\mathbf{V}^{t+1} = \mathbf{V}^*$
    $t = t + 1$
until $\|\mathbf{W}^t - \mathbf{W}^{t-1}\|_2 + \|\mathbf{U}^t - \mathbf{U}^{t-1}\|_2 + \|\mathbf{V}^t - \mathbf{V}^{t-1}\|_2 < \epsilon$
```

the parameter matrices are as following:

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{U}} &= ((\mathbf{WX}) \circ (\mathbf{V}(g(\mathbf{V}^T((\mathbf{UY}) \circ (\mathbf{WX})) - \mathbf{A}))\mathbf{Y}^T + \lambda_1\mathbf{U} \\ \frac{\partial J}{\partial \mathbf{V}} &= ((\mathbf{WX}) \circ (\mathbf{UY}))(g(\mathbf{V}^T((\mathbf{UY}) \circ (\mathbf{WX})) - \mathbf{A})^T + \lambda_2\mathbf{V} \\ \frac{\partial J}{\partial \mathbf{W}} &= ((\mathbf{UY}) \circ (\mathbf{V}(g(\mathbf{V}^T((\mathbf{UY}) \circ (\mathbf{WX})) - \mathbf{A}))\mathbf{X}^T + \lambda_3\mathbf{W}\end{aligned} \tag{12}$$

Here, $\circ$ denotes the Hadamard product. With two of the parameter matrices fixed, we need to estimate the optimal value of the third matrix according to one of these equations using L-BFGS algorithm.

3.3. Enhancing Category Information

As fine-grained semantic descriptions, attributes are usually hard to define and costly to acquire. Therefore, the scale of labeled attribute dataset is relatively small compared to those in large-scale visual learning tasks, such as image classification and image search. To address the small scale training data problem, our method gives a way to boost attribute learning by enhancing category information.

Assuming there are two types of training data $\mathbf{X}$ and $\mathbf{X}_a$. The former has both attribute labels and category labels while the latter only has category labels $\mathbf{Y}_a$. The objective function of our multiplicative framework can be written as:

$$\begin{aligned}\tilde{J}_a \quad = \quad & L(\mathbf{X}, \tilde{\mathbf{Y}}, \mathbf{A}; \Phi) + \beta L_Y(\mathbf{X}, \mathbf{X}_a, \mathbf{Y}, \mathbf{Y}_a; \Theta) \\ & + \lambda_1\Omega(\mathbf{W}) + \lambda_2\Omega(\mathbf{U}) + \lambda_3\Omega(\mathbf{V}) + \lambda_4\Omega(\Theta).\end{aligned} \tag{13}$$

With the enriched training data, we can obtain enhanced object category information $\tilde{\mathbf{Y}}$, which benefits attribute learning.

4. Experiments

4.1. Datasets and Evaluations

To access the efficacy of our proposed multiplicative framework, we conduct experiments on real-world datasets for attribute prediction and zero-shot learning. Two types of attribute definition are adopted in our experiments. For category-level attribute definition, we use Animals with Attributes and Caltech-UCSD Birds. These two datasets are widely used to verify the transferability of the learned attribute classifiers. For instance-level attribute definition, aPascal-aYahoo and ImageNet attributes are used to validate the discriminative power of the proposed methods. The detailed information of the above four datasets are as follows:

Animals with Attributes (AwA) [17]. The dataset is collected by querying the image search engines with images from 50 animal categories. Outliers and duplicates are further removed manually, and the number of the remaining images is 30,475. The minimum and maximum number of images from one category is 92 and 1,168 respectively. Each category is annotated with 85 attribute labels.

aPascal-aYahoo (aPaY) [8]. The first part of this dataset is called aPascal which contains 6430 training images and 6355 testing images from Pascal VOC 2008 challenge. Each image comes from twenty object categories. The second part is aYahoo dataset. There are 2644 images belonging to twelve categories which are disjoint with aPascal dataset. Each image is annotated with 64 binary attribute labels in these two datasets. In our experiment, we merge them into one whole dataset.

Caltech-UCSD Birds (CUB) [19]. This dataset contains 11,788 images of 200 bird classes. Each category is annotated with 312 attributes. Since this dataset gives a fine-grained category description, it seems harder to leverage categroy information to promote attribute learning than AWA.

ImageNet Attributes (INA) [18]. ImageNet Attribute dataset contains 9,600 images from 384 categories. Each image is annotated with 25 attributes describing color, patterns, shape and texture. 3-4 workers are asked to provide a binary label indicating whether the object in the image contains the attribute or not. When there is no consensus among the workers, the attribute will be labeled as ambiguous for this image.

For attribute prediction, we randomly split the datasets into three subsets with equal size for training, validating and testing. The dimension of latent space is set to the minimum of the number of categories and attributes. Other parameters are tuned on the validation set. We use the 4096-D DeCAF features which are extracted by the Convolutional Neural Networks (CNN) described in [7]. The performance of attribute predictors are measured by *mean area under ROC* *curve* (mAUC) and *mean classification accuracy* (mACC).

For Zero-shot learning, we use the specified seen and unseen class splits of AwA. The seen part contains 24,295 images of 40 classes and the remaining 6,180 images from 10 classes are used as unseen data. Considering the CUB dataset does not provide the specific seen and unseen class splits, we use 150 classes as seen classes and leave the remaining 50 classes as unseen. For AwA dataset, we use the DeCAF features provided on the dataset website [1] . As for CUB, we extract 4096-D DeCAF features using the method in [7]. In these experiments, we randomly choose ten percent of the seen data for validation to tune the parameters. The dimension of latent space is set to the minimum of the number of seen categories and attributes. The performance of zero-shot learning are evaluated by *normalized multi-class accuracy*.

4.2. Category-level Attribute Prediction

For category-level attribute prediction, we choose AwA dataset to compare the propose method with some related methods including:

1. Direct attribute prediction (DAP). We use linear SVMs to train attribute classifiers separately. The Lagrangian multipliers of SVMs are set to the same value for all attribute classifiers without tuning elaborately.

2. Indirect attribute prediction (IAP). We train one-versus-rest linear SVMs to do multi-class classification. The probability of class prediction is acquired by using LIBSVM package[4].

3. Multi-task attribute prediction (MTAP). Besides DAP and IAP, we also compare our method with multi-task attribute learning method [12]. Besides the conventional lasso across attribute groups, they also apply the ℓ_{21} regularizer for within-group sharing.

4. Direct concatenation method (Concat.). We simply concatenates the image and class prediction, $\mathbf{X}$ and $\tilde{\mathbf{Y}}$, as the input for attribute learning. This method is a strong baseline, since it leverages the auxiliary information in an additive way.

5. The proposed unified multiplicative framework(UMF-IS) with instance-specific attribute classifiers.

In this experiment, the dimension of the latent common space is set to 50. β is set to 0.5 and $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are all set to 10^{-5}.

From the results summarized in Table 1, we can see that our UMF model achieves the highest mACC and comparative mAUC with IAP. The reason why IAP achieves a high mAUC but a relatively poor mACC is that the unbalance of positive and negative samples of attribute labels does not affect IAP model, which predicts attributes only from object categories. The performance of MTAP is not as promising as being expected. From our perspective, the feature we used is highly sparse because the rectified linear units of CNN discard the negative part of the input. As a result, the power of sparse regularization decreases tremendously. In addition, the features extracted from a higher CNN layer represent abstract visual patterns instead of low-level patterns, such as color or shape.

Table 1. Comparison of different attribute learning methods on AwA dataset

Methods	mAUC	mACC(%)	Aux. Info
DAP	0.899	86.3	No
IAP	**0.927**	86.9	Yes
MTAP	0.903	86.6	Yes
Concat.	0.913	87.3	Yes
UMF-IS	0.926	**89.7**	Yes

4.3. Instance-level Attribute Prediction

4.3.1 Enhancing Instance-level Attribute Prediction

In this experiment, we compare our method UMF-IS with traditional instance-level attribute prediction method DAP on INA dataset. For this purpose, we train our model by using images from the training set. The dimension of the latent common space is set to 25 and β is set to 0.5. $\lambda_1, \lambda_2, \lambda_3$ are all set to 10^{-4} and λ_4 is set to 10^{-5}. Then, we validate the effectiveness of enhancing attribute prediction ability by leveraging auxiliary category information (named as UMF-IS-Aux). To this end, we fix the above parameters and use Eqn. (14) to retrain the model by considering validation set as auxiliary data. One thing to note is only category labels of validation set are used.

Table 2. Instance-level attribute prediction on INA dataset

Methods	mAUC	mACC(%)
DAP	0.899	82.7
UMF-IS	0.918	83.2
UMF-IS-Aux	0.931	83.8

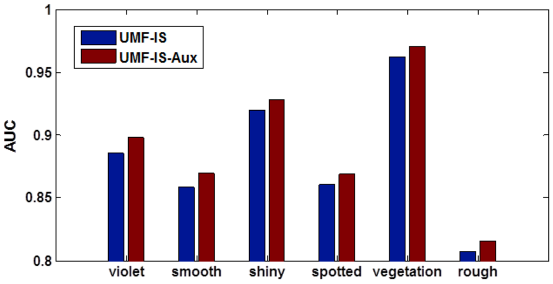


Figure 3. Attributes that are enhanced most with auxiliary data

The experimental results are shown in Table 2. Among all the 25 attributes in INA dataset, the attribute prediction

[1] http://attributes.kyb.tuebingen.mpg.de/

performance of 22 attribute are improved. For the left three attributes white, wooden and yellow, the attribute prediction performance is decreased slightly. We show six attribute prediction which are enhanced most in Fig.3. Since the annotation of instance-level attribute is costly, enhancing attribute prediction by using auxiliary task such as object recognition seems promising.

4.3.2 Category-Sensitive Attribute Prediction

In real-world application, some attributes have different visual appearances across categories. We call them category-sensitive attributes. For example, the rectangular property of a comb varies from that of a window based on visual information. Assuming we know the category information of an image, we should predict the presence or absence of attributes using its own category-specific attribute classifiers. In this experiment, we compare our category-specific (UMF-CS) and instance-specific (UMF-IS) methods with related category-sensitive attribute classifiers which are trained for each category-attribute pair. In the experiments, only the category-attribute pair which has both positive and negative image exemplars are valid for training a classifier. The valid category-attribute pairs are shown in Figure 4. We count the number of valid pairs and report the statistical results in Table 3.

The methods involved in the comparative study include:

1. Universal attribute prediction (U). A linear SVM classifier is trained to predict one binary attribute universally. In such case, the positive images are irrelevant to their categories.

2. Category-sensitive attribute prediction (CS). Instead of learning attribute classifier without considering category information, an importance-weighted linear support vector machine is used to predict attribute category-dependently. To train the t^{th} attribute classifier for category j, the violating attribute label constraints for positive and negative samples from category j are given a higher penalty, as suggested in [5].

During training, we train importance-weighted attribute classifiers on 383 and 607 valid category-attribute pairs for aPaY and INA respectively. For our methods, both the category-sensitive and instance-sensitive attribute classifiers are trained jointly without considering the instance weight. In INA dataset, some attribute labels are ambiguous, we simply set them to 0.5 for joint training.

For test stage, we do attribute prediction for all the valid category-attribute pairs in the test set. The category of the test image is assumed to be known in this setting. At last, we average the AUCs of all the valid category-attribute pairs to show the effectiveness of the comparative methods.

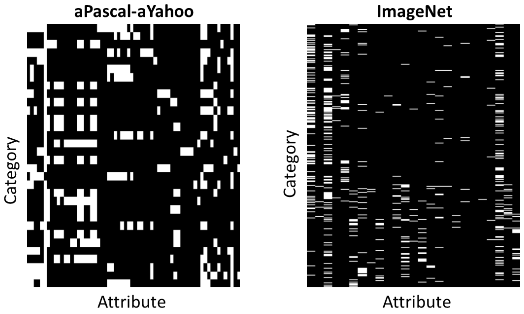


Figure 4. Valid category-attribute pairs. The white entities denote the category-attribtue pairs have both positive and negative exemplars.

As shown in Table 3, category-sensitive attribute classifier achieves much better performance than the universal attribute classifier. However, the number of trained classifiers are proportional to $C \times T$. Therefore, when the dataset has a large amount of attributes and categories, it is costly to train all of them individually. Moreover, the relationship between attribute and attribute is totally lost. For our methods, UMF-CS and UMF-IS both achieve better performance than universal attribute classifiers and outperform category-sensitive attribute classifier on INA dataset. At the same time, the number of classifiers are $C + T$ for UMF-CS and UMF-IS. Therefore the scalability of our methods is guaranteed.

4.4. Zero-shot Learning Based on Attributes

Since different images may share common attributes, we can recognize images from unseen classes based on transferred attribute concepts, which is referred as as *zero-shot learning* [17]. Traditional multi-class classifiers can not tackle this task since no training data is available to learn the parameters. We assume there are K seen classes $\{y_1, y_2, \cdots, y_K\}$ and L unseen classes $\{z_1, z_2, \cdots, z_L\}$. The attribute classifiers are learned based on the K seen classes. During testing, the unseen category of an image $\mathbf{x}$ is determined based on the posterior probability computed by leveraging on the known attribute-category relations:

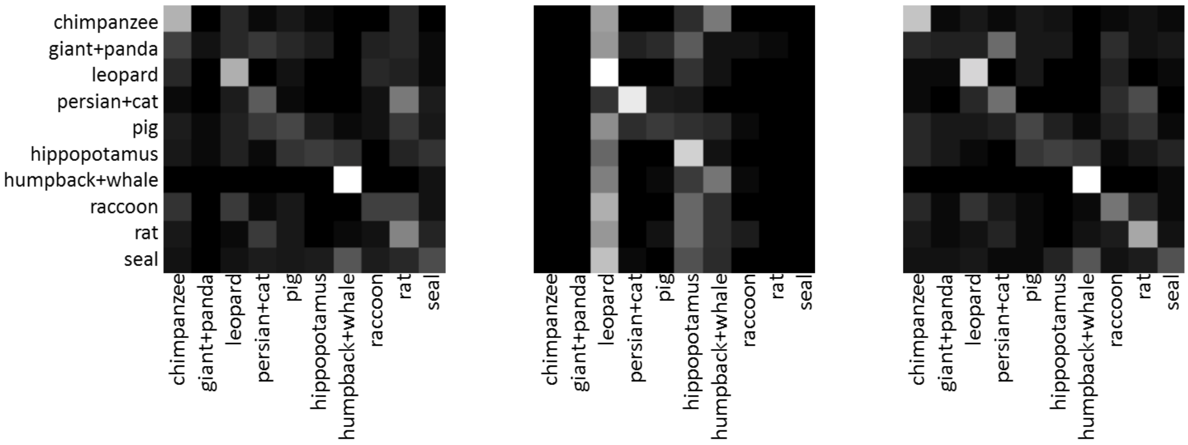
$$P(z_l|\mathbf{x}) = \frac{p(z_l)}{p(a^{z_l})} \prod_{t=1}^{T} p(a_t^{z_l}|\mathbf{x}), \tag{14}$$

where $a_t^{z_l}$ is the t-th attribute label of class z_l. Based on simple assumption, the class prior $p(z_l)$ is identical for all the classes and $p(a^{z_l})$ is assumed to be a factorial distribution $p(a^{z_l}) = \prod_{t=1}^{T} p(a_t^{z_l})$.Based on the seen classes, the attribute priors are defined as $p(a_t) = \frac{1}{K}\sum_{k=1}^{K} a_t^{y_k}$. For our method, the attribute predictive probability has the following form:

$$p(a_t^{z_l}|\mathbf{x}, \tilde{\mathbf{y}}) = \frac{[\![a_t^{z_l} = 1]\!]}{1 + e^{-f(\mathbf{x},\tilde{\mathbf{y}},t)}} + \frac{[\![a_t^{z_l} = 0]\!]}{1 + e^{f(\mathbf{x},\tilde{\mathbf{y}},t)}} \tag{15}$$

Table 3. Attribute prediction on category-attribute pairs

	Dataset Info		# Classifiers			mean AUC			
	Categ	attr	U	CS	UMF	U	CS	UMF-CS	UMF-IS
aPascal-aYahoo	32	64	64	383	96	0.584	0.618	0.600	0.603
Imagenet	384	25	25	607	409	0.770	0.843	0.841	0.847



(a) DAP (b) IAP (c) Our method

Figure 5. Confusion matrices between 10 unseen classes on AwA. The rows are the truths and the columns are the predictions.

Table 4. Comparison of different zero-shot learning methods. Feature type H, FV and D represent hand-crafted, fisher vector and DeCAF. '*' means our implementation. '–' means no results.

Methods	AWA	CUB	Fea Type
DAP	41.4[17]/45.3*	–/16.9*	H/D
IAP	42.2[17]/46.4*	–/16.7*	H/D
ALE	37.4[2]/45.7[1]	18.0[2]/**20.2**[1]	FV/D
BN	43.4[21]	–	H
TMV-HLP	47.1[9]	–	H
HAP-H	45.0[10]	17.5[10]	D
HAP-G	45.0[10]	17.5[10]	D
UMF-IS w/o $\mathbf{W}$	43.7	15.9	D
UMF-CS	47.1	17.7	D
UMF-IS	**48.6**	18.2	D

where $[\![\cdot]\!]$ is the Iverson's bracket notation. The form of function f is defined in Eqn. (1).

In the experiments, we compare our methods with DAP, IAP, Bayesian network (BN) [21], transductive multi-view Bayesian label propagation (TMV-BLP) [9] and Hypergraph-regularized Attribute Predictors (HAP) [10]. We also validate the influence of $\mathbf{W}$ by introduce a variant of our method (UMF-IS w/o $\mathbf{W}$) which means using multiplicative model on the original input feature space. The dimension of the latent common space is set to 40 and 150 for AwA and CUB respectively.

As shown in Table 4, the performance of our method is significantly better than the state-of-the-art approaches. In addition, the accuracy of UMF w/o $\mathbf{W}$ is lower than UMF by about 5 percentage, showing that the latent common space learned by $\mathbf{W}$ is essential for our multiplicative model to disentangle the factors and learn the intrinsic property of attributes. From Fig. 5, we can also see the superiority of our method. Especially, our method can distinguish "raccoon" from "rat" much more clearly than DAP.

5. Conclusions

We introduce a unified multiplicative framework for attribute learning by leveraging category information. Unlike the methods which predict attributes only based on the visual appearances, our model explicitly captures the relationship among image, attribute and category in a multiplicative way in the latent feature space. We perform experiments on four widely used datasets for attribute prediction and zero-shot learning. The empirical results show that our method achieves better performance in attribute prediction on public datasets with whether instance-level or category-level annotation. In addition, the proposed model can be enhanced by auxiliary data, which reduces the effort of instance-level attribute annotation to some extent. Moreover, our method significantly improves the accuracy of zero-shot learning, verifying that the attribute classifiers learned by our method have better generalization ability.

6. Acknowledgements

This work is partially supported by 973 Program under contract No. 2015CB351802, and Natural Science Foundation of China under contracts Nos. 61390515, 61272319, and 61222211.

References

[1] Z. Akata, H. Lee, and B. Schiele. Zero-shot learning with structured embeddings. *arXiv preprint arXiv:1409.8403*, 2014. 2, 8

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 819–826. IEEE, 2013. 2, 8

[3] J. Bi, T. Xiong, S. Yu, M. Dundar, and R. B. Rao. An improved multi-task learning approach with applications in medical diagnosis. In *Machine Learning and Knowledge Discovery in Databases*, pages 117–132. Springer, 2008. 4

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. 6

[5] C.-Y. Chen and K. Grauman. Inferring analogous attributes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 200–207. IEEE, 2014. 3, 7

[6] L. Chen, Q. Zhang, and B. Li. Predicting multiple attributes via relative multi-task learning. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1027–1034. IEEE, 2014. 1, 2

[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of The 31st International Conference on Machine Learning*, pages 647–655, 2014. 5, 6

[8] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009. 1, 5

[9] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *Computer Vision–ECCV 2014*, pages 584–599. Springer, 2014. 2, 8

[10] S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang. Learning hypergraph-regularized attribute predictors. *arXiv preprint arXiv:1503.05782*, 2015. 2, 3, 8

[11] S. J. Hwang and L. Sigal. A unified semantic embedding: Relating taxonomies and attributes. In *Advances in Neural Information Processing Systems*, pages 271–279, 2014. 2

[12] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1629–1636. IEEE, 2014. 2, 6

[13] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 595–603, 2014. 3

[14] R. Kiros, R. Zemel, and R. R. Salakhutdinov. A multiplicative model for learning distributed text-based attribute representations. In *Advances in Neural Information Processing Systems*, pages 2348–2356, 2014. 3

[15] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2973–2980. IEEE, 2012. 1

[16] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009. 1

[17] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(3):453–465, 2014. 1, 3, 5, 7, 8

[18] O. Russakovsky and L. Fei-Fei. Attribute learning in large-scale datasets. In *European Conference of Computer Vision (ECCV), International Workshop on Parts and Attributes*, 2010. 5

[19] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011. 5

[20] X. Wang, J. Bi, S. Yu, and J. Sun. On multiplicative multitask feature learning. In *Advances in Neural Information Processing Systems*, pages 2411–2419, 2014. 2

[21] X. Wang and Q. Ji. A unified probabilistic approach modeling relationships between attributes and objects. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2120–2127. IEEE, 2013. 2, 3, 8

[22] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *Computer Vision–ECCV 2010*, pages 155–168. Springer, 2010. 1