This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2836461, IEEE Transactions on Pattern Analysis and Machine Intelligence
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2018.2836461

Unifying Visual Attribute Learning with Object Recognition in a Multiplicative Framework

Kongming Liang, *Student Member, IEEE,* Hong Chang, *Member, IEEE,* Bingpeng Ma,
Shiguang Shan, *Senior Member, IEEE,* Xilin Chen, *Fellow, IEEE*

Abstract—Attributes are mid-level semantic properties of objects. Recent research has shown that visual attributes can benefit many typical 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 to fulfil attribute prediction. The resulting attribute classifier is category-specific instead of being shared by all categories. Moreover, our model can leverage auxiliary data to enhance the predictive ability of attribute classifiers, which can reduce the effort of instance-level attribute annotation to some extent. By integrated into an existing deep learning framework, our model can both accurately predict attributes and learn efficient image representations. Experimental results show that our method achieves superior performance on both instance-level and category-level attribute prediction. For zero-shot learning based on visual attributes and human-object interaction recognition, our method can improve the state-of-the-art performance on several widely used datasets.

Index Terms—Attribute learning, zero-shot learning, image understanding.

1 Introduction

Attributes are namable properties of objects which are observable from visual images. In computer vision community, two types of attribute are defined according to the level of annotation which are named as instance-level attributes and category-level attributes. By annotating each visual instance separately, instance-level attributes can provide semantic descriptions, such as the holistic perception (e.g., color, shape, etc.) and presence or absence of local parts for images. They have been successfully applied in describing and understanding general objects [1], [2], [3] and even some specific objects such as birds [4] and human faces [5], [6]. As a mid-level semantic cue, they can bridge the gap between low-level features and high-level categorization. Recent research has verified that instance-level attributes can benefit many traditional learning problems (e.g., image search [7], object recognition [8] and face verification [5]). For category-level attributes, images belong to the same category share common attribute annotations. The presence or absence of category-level attributes are judged based on human subjects on the ”relative strength of association” between attributes and categories [9], [10]. By transferring from seen classes to unseen classes, they provide a proper way to address *zero-shot classification* [11] which is significant in alleviating prohibitive data collection and annotation for large scale object recognition.

Direct attribute prediction methods [1], [11] are the most widely used methods for attribute learning problems. They train a group of binary classifiers from image-attribute pairs, one individually for each attribute. Fig. 1 (a) illustrates the direct attribute prediction method. During test stage, the learned classifiers are applied to predict which subset of attributes the input image may have. In addition, they can further exploit knowledge transfer to tackle zero-shot classification. 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 due to the individual training strategy. Naturally, visual attributes as properties of objects are correlated with each other, therefore it is more appropriate to learn all the attributes jointly, through sharing attribute-specific parameters or common semantic representations.
2) Some attributes are hard or even unable to predict based on visual appearances due to the lack of relevant information. For example, it is impossible to infer color-relevant attribute from an gray image 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 [11], 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) Distraction from object interaction. In real-world scenario, objects from different categories usually

- *The authors are with Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China and University of Chinese Academy of Sciences, Beijing 100049, China.*
E-mail: {kongming.liang, bingpeng.ma}@vipl.ict.ac.cn, {changhong, sgshan, xlchen}@ict.ac.cn

0162-8828 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Copyright (c) 2018 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2836461, IEEE Transactions on Pattern Analysis and Machine Intelligence
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2018.2836461

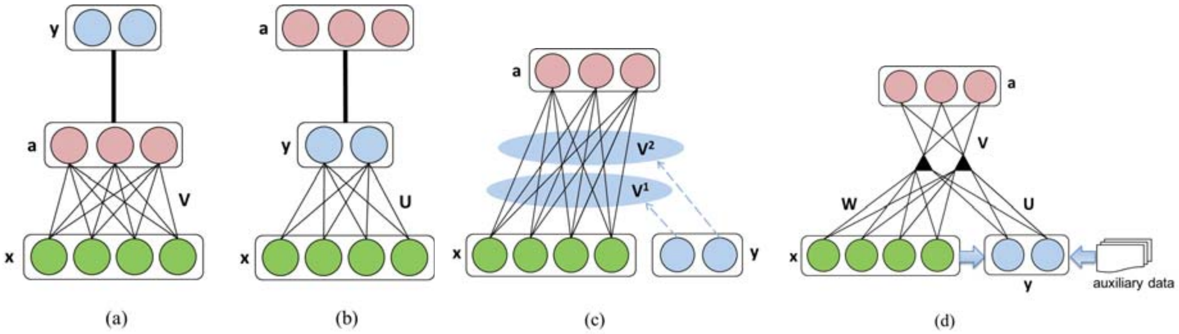


Fig. 1. Models for attribute learning. (a) direct attribute prediction model; (b) indirect attribute prediction model; (c) category-sensitive attribute learning model; (d) our proposed multiplicative model. $\mathbf{x}, \mathbf{a}$ and $\mathbf{y}$ denote image, attribute and label vectors respectively.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

interact with each other, such as a man is riding a horse, a dog is sitting on a sofa. As a result, the attribute of the centric object may be distracted by the other overlapped objects.

5) Different visual attribute manifestations vary across categories. Humans usually describe visual attribute of different categories by using the same semantic word. However, visual attribute across categories appears in relatively large change. For example, the same attribute concept "fluffy" varies considerably between dog and towel [12].

Some methods have been proposed to tackle the above problems to some extent. Lampert et al. [11] propose a method to indirectly predict attributes by transferring 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. [13] and Chen et al. [12] formulate attribute learning in regularization-based multi-task learning framework, where each subtask corresponds to learning one attribute. Wang et al. [14] use Bayesian network to enhance attribute prediction by leveraging the statistical relationships between attributes and objects. Huang et al. [15] 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 (UMF) to address the aforementioned problems. Fig.1 (d) illustrates our model, where the image and category vector 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. The common space is introduced for two reasons. Firstly, it facilitates representation learning for all the attributes by sharing the projection matrix of the input image feature. Secondly, category information can be elegantly integrated into the original input feature for enhancing attribute prediction. In this way, the attribute classifiers for input images lead to be category-specific and can efficiently model the third-order relationship among image, category and attribute. However, under most circumstance, category label is not provided when we do attribute prediction. Therefore we integrate multi-class categorization into the original unified multiplicative multi-task learning framework. The input category information is replaced as the category prediction from a multi-class classifier instead of the ground truth category label. Consequently, the resulting attribute classifiers are instance-specific (a linear combination of the category-specific attribute classifiers). A weakness of the proposed method is the discrepancy between training and testing category prediction which is under a similar scenario for sequence generation [16]. To tackle the above problem, we present a variant of the proposed method by making the attribute prediction conditioning on the hidden representations of multi-class categorization task. Therefore, the input of the proposed model is more consistent among training and testing stages. Experimental results show that the new variant of the proposed method achieves better performance on attribute prediction.

In recent years, deep learning has drawn much attention in both computer vision and machine learning communities. By learning representation of objects, convolutional neural network has dominated the performance on object recognition and detection [17], [18], [19]. Moreover, it also brings about breakthroughs in many relevant tasks such as face recognition [20], fine-grained classification [21] and semantic segmentation [22]. Some previous works [23], [24], [25] show that the the off-the-shelf deep learning feature (e.g. DeCAF [24], VGG [25]) can be repurposed to novel generic tasks. However, one of the factors of success for deep learning comes from the end-to-end model training manner. An advantage of our method is it can be easily integrated into deep learning framework and optimise attribute prediction joint with representation learning. In this way, the learned representation are more suitable for attribute prediction task.

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 classifiers in our method are instance-specific and can be decomposed into a linear combination of category-specific attribute classifiers.

0162-8828 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Copyright (c) 2018 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2836461, IEEE Transactions on Pattern Analysis and Machine Intelligence
This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2018.2836461

Thus, it gives a finer attribute description than conventional attribute classifiers and can be transferred to unseen class more easily. (4) Our method can be integrated into the off-the-shelf deep learning framework. Therefore, the learned representation are more powerful for attribute prediction. (5) 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 and its variant in detail. Experimental results are then shown in Section 5.1. Finally, Section 6 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. [26], [27] 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. [28] 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. [29] propose a framework called transductive multi-view embedding to tackle the projection domain shift problem in zero-shot learning by integrating multiple types of side information.

Multi-task learning. Multi-task learning is intended to improve generalization performance by joint training multiple tasks. Various types of multi-task learning methods have been proposed to learn the intrinsic correlation between tasks by using feature sharing and parameter regularization strategies. Attribute learning problem can be formulated in the multi-task learning framework, where each task corresponds to learning one semantic attribute. Jayraman et al. [13] propose to decorrelate visual attributes by developing a 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. [12] integrate relative attribute learning with a robust regularization scheme. By joint training all the attribute together, the proposed method can detect outlier attribute and leverage correlations among attributes simultaneously.

Multiplicative models. Multiplicative models are effective in relating separate underlying factors in data. [30] 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 among image, category and attribute. And the cross-task component in our method is not under the assumption of being a vector. [31] 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 [31], which makes use a known language model. [32] 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. [11] 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. Hwang et al. [33] propose to learn a shared lower dimensional representation by optimizing a joint loss function with both attribute and object tasks. The shared representations are learned with the common sparsity patterns across both types of prediction tasks. Instead of tackling the two tasks symmetrically, our proposed method considers object categorization as an auxiliary task. Wang et al. [14] propose a unified probabilistic model to capture the class-dependent and class-independent attribute relationships, which benefit both attribute prediction and object recognition. [34] 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. [15] 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. Gan et al. [35] treat each category as a source domain and learn attribute detectors which can well generalize across different categories with multi-source domain learning.

Zero-shot Learning. Existing object recognition system starves for training data which need manual labelling for each category. Since the world contains tens of thousands of different object classes, it is costly and even impossible to annotate a large collection of images. To tackle the above problem, zero-shot learning is presented to utilize knowledge transferred from seen classes to recognise unseen classes. Attributes as mid-level semantic representation are widely used in previous methods to facilitate zero-shot learning. In addition, Ba et al. [36] propose to mining text information to acquire semantic relationship between categories. Similarly, Akata et al. [27] consider semantic class taxonomies as prior information. Beyond that, Fu et al. [29], [37] exploit multiple semantic information which are complementary with each other to improve the conventional approaches. Except for leverage multiple types of side information, some existing problems for zero-shot learning are well explored. Jayraman et al. [38] propose a method based on random

0162-8828 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Copyright (c) 2018 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2836461, IEEE Transactions on Pattern Analysis and Machine Intelligence
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2018.2836461

forest to tackle unreliability of attribute prediction on novel classes. Jayraman et al. [13] exploit a multi-task regularizer to decorrelate attributes form different types. Fu et al. [29] propose a transductive framework to tackle the domain shift problem between seen categories and unseen categories.

Deep Attribute Learning. Deep learning has improved the state-of-the-art performance on various tasks in computer vision. For well estimating millions of parameters of deep learning framework, a large scale supervision is needed. Comparing with generic object recognition, attributes need to be annotated with multiple labels for each image instead of a single category label. Therefore, datasets with attribute annotation are relatively small and it is not optimal to train a deep learning model from scratch to do attribute prediction. Shankar et al. [39] propose a novel training algorithm for convolutional neural network called Deep-Carving to predict multiple attributes with only partial labels. Zhang et al. [40] present a method to predict human attributes by combining part-based models and deep learning to tackle pose and viewpoint variation. Liu et al. [6] propose to use two cascaded convolutional neural networks which are pretrained with general objects and face identities respectively. And the first net is used to do face detection while the second net facilitates facial attribute prediction. In this paper, we integrate the proposed multiplicative multi-task learning model with convolutional neural network pretrained on object recognition task.

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.

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:

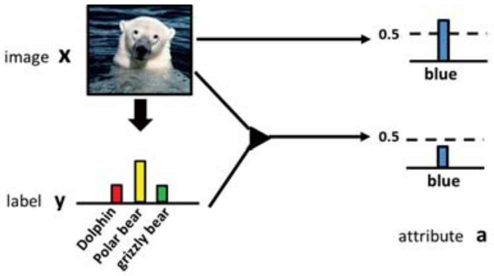


Fig. 2. Illustration of our method for predicting attribute “blue”. Category information helps to address the negative correlation problem. 0.5 denotes the decision boundary for attribute prediction.

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

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. For each attribute, the input feature is in a bilinear form [41], which characterizes the interaction between the two input factors (image and category). 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, category information is an important factor for attribute learning, as the visual appearances of attribute usually vary across categories. Besides, for some attributes which are hard to predict based on visual cues, we can infer them from category information. Moreover, category information may be helpful to address the negative correlation problem, as illustrated in Fig. 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 and category information for better attribute prediction.

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:

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

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:

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

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 $\Omega(\cdot)$.

0162-8828 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Copyright (c) 2018 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2836461, IEEE Transactions on Pattern Analysis and Machine Intelligence
This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2018.2836461

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:

$$\begin{aligned} f(\mathbf{x}_i, \mathbf{y}_i, t) &= ((\mathbf{U}\mathbf{y}_i) \odot \mathbf{v}_t)^T (\mathbf{W}\mathbf{x}_i) & (5) \\ &= \left(\sum_{j=1}^{C} y_{ji} (\mathbf{u}_j \odot \mathbf{v}_t) \right)^T (\mathbf{W}\mathbf{x}_i). & (6) \end{aligned}$$

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.

In the test stage, the category of an image is usually not provided for attribute prediction. To construct category-specific attribute classifier, we first build a softmax multi-class object classifier by minimizing the following regularized loss:

$$\begin{aligned} & L_Y(\mathbf{X}, \mathbf{Y}; \Theta) + \lambda_4 \Omega(\Theta) \\ = \; & \frac{1}{N} \sum_{j=1}^{C} \sum_{i=1}^{N} -1\{y_{ji} = 1\} \frac{exp(\boldsymbol{\theta}_j^T \mathbf{x}_i)}{\sum_{k=1}^{C} exp(\boldsymbol{\theta}_k^T \mathbf{x}_i)} \\ & + \lambda_4 \Omega(\Theta) \qquad (7) \end{aligned}$$

where θ_j is the classifier parameter for object category j. λ_4 is the weight decay term to tackle the overfitting problem. The parameters Θ are estimated from the training data. Then, the category probability of a test image $\mathbf{x}$ can be estimated as follows:

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

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

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

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 object classification 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:

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

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

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

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.

3.1.3 Discussion

The main difference between the category-specific and instance-specific attribute classifiers lie in the time and scheme of combining category information into the unified framework. For instance-specific scheme, the category information is leveraged before training the attribute classifier. Therefore, we call it early fusion. For category-specific scheme, we first train attribute classifier for all categories and marginalize them at the last stage. So it can be considered as a late fusion scheme.

Actually, this topic has also been studied in previous work. Snoek et al. [42] validate the two schemes in semantic video analysis. Dong et al. [43] propose a performance evaluation of early and late fusion schemes for semantics indexing. In addition, fusion scheme can benefit many other tasks, such as object recognition [44], [45], [46], biometric analysis [47], video event detection [46], [48]. However, the superiority of these two schemes is not consistent. Compared with late fusion approaches, early fusion shows more significance in statistic for semantics indexing. As for semantic video analysis, late fusion tends to give slightly better performance for most concepts. For attribute prediction, we can empirically prove that instance-specific classifiers are superior to category-specific classifiers.

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. The details are illustrated in Fig.3.

3.2 Optimization

A multiplicative models can be 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 [49]. 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 the parameter matrices are as following:

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

0162-8828 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Copyright (c) 2018 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2836461, IEEE Transactions on Pattern Analysis and Machine Intelligence
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2018.2836461

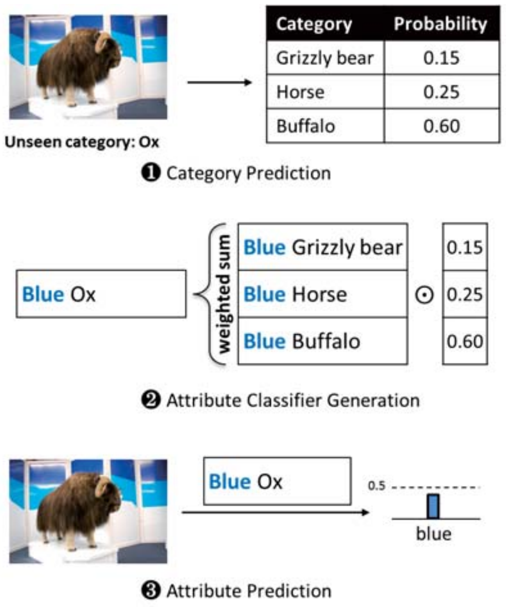


Fig. 3. Pipeline of attribute prediction for unseen categories.

```
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 $\|\mathbf{W}^t - \mathbf{W}^{t-1}\|_2 + \|\mathbf{U}^t - \mathbf{U}^{t-1}\|_2 + \|\mathbf{V}^t - \mathbf{V}^{t-1}\|_2 < \epsilon$
```

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. For zero-shot learning, we fix the weights of $\mathbf{W}$ and $\mathbf{V}$ and only optimize the category-relevant term $\mathbf{U}$ to increase the generalization ability between seen and unseen categories.

4 EXTENSION OF THE PROPOSED METHOD

4.1 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:

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

The parameters used to predict categorization are optimized with the enriched training data. Therefore, we can obtain more accurate object category information $\tilde{\mathbf{Y}}$ to benefits attribute learning.

4.2 Tackling unreliable category information

A common problem for category-sensitive and instance-sensitive attribute classifiers is that they both ignore the unreliable category information during the test stage. For category-sensitive attribute classifier, the ground truth category label is only provided during training stage which assumes the category label should always be right. This problem also exists for instance-sensitive attribute classifiers.

A solution to tackle the above problem is to avoid the attribute classifier directly conditional on the category label or prediction. That means we should leverage another form of category information whose distribution is relatively consistent on training and testing images. To achieve this goal, we propose to make the attribute classifier conditioned on the hidden representation for categorization since the hidden units have more generalization ability than the high-level category prediction. We first learn hidden representations of images, which are then used to predict category label or further multiplied with the direct input feature for attribute prediction. The hidden units for the i^{th} image are denoted as $\mathbf{h}_i = g(\mathbf{U}'\mathbf{x}_i)$ where $\mathbf{U}'$ is the projection parameter. The hidden units of training images are expressed in a matrix form as $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N] \in \mathbb{R}^{F\times N}$.

The discriminant function of the t^{th} attribute of the object in image $\mathbf{x}_i$ is then defined as follows:

$$f_H(\mathbf{x}_i, \mathbf{y}_i, t) = \langle \mathbf{v}_t, \mathbf{h}_i, g(\mathbf{W}\mathbf{x}_i)\rangle. \tag{14}$$

The non-linear function $g(\mathbf{W}\mathbf{x}_i)$ is used to make the input image feature to be symmetric with the hidden feature $\mathbf{h}_i$. The objective function of our multiplicative model based on the hidden units takes the following form:

$$\begin{aligned}\tilde{J}_H \quad = \quad & L_H(\mathbf{X}, \mathbf{H}, \mathbf{A}; \Phi') + \beta L_Y(\mathbf{H}, \mathbf{Y}; \Theta) + \lambda_1\Omega(\mathbf{W}) \\ & + \lambda_2\Omega(\mathbf{U}') + \lambda_3\Omega(\mathbf{V}) + \lambda_4\Omega(\Theta),\end{aligned} \tag{15}$$

where $\Phi' = \{\mathbf{W}, \mathbf{U}', \mathbf{V}\}$. $L_H(\cdot)$ is in a similar form as the Eqn. (3) by replacing $f(\cdot)$ with $f_H(\cdot)$. From the above equation, we can see the hidden units are used for both category and attribute prediction. This variant unified multiplicative framework with hidden representation learning is called UMF-H, as illustrated in the right part of Fig. 4. Except for tackling uncertainty of the category information, another benefit of UMF-H is that the initialization of $\mathbf{W}, \mathbf{V}$ does not rely on the SVD decomposition which is a must for the original method as depicted in Algorithm 1. So we can

0162-8828 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Copyright (c) 2018 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2836461, IEEE Transactions on Pattern Analysis and Machine Intelligence
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2018.2836461

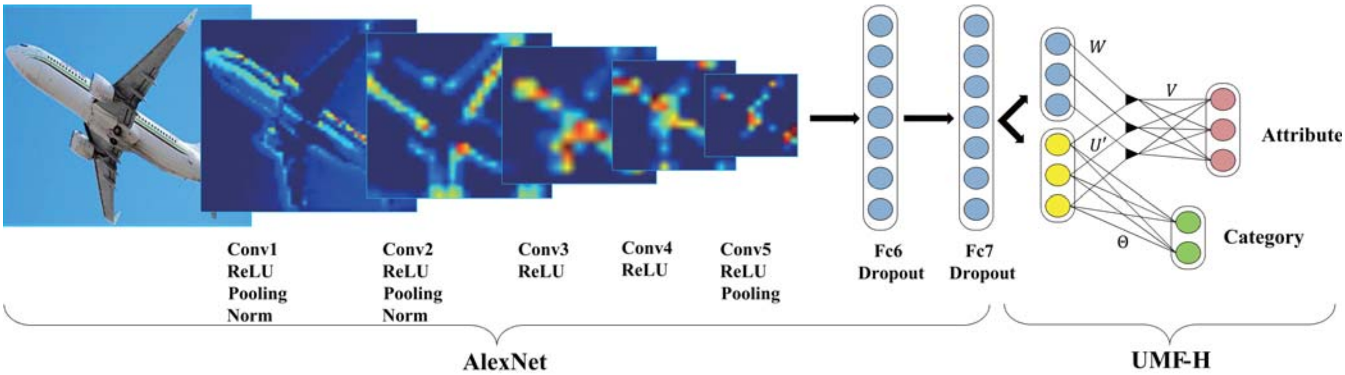


Fig. 4. UMF-Deep model integrating AlexNet with UMF-H. The black arrow indicates fully connection. The yellow circles are hidden units which are used for both attribute prediction and categorization.

initialize them randomly in a similar way of initializing **U**. The experimental results show that leveraging the hidden representation is essential for further improvement over the origin formulation.

4.3 Integration with deep learning framework

The proposed UMF-H can be easily stacked into existing deep learning framework. In this paper, we use the Alexnet as our basic model. It has five convolutional layers and two fully-connected layers with dropout. Between some of these layers, ReLU layers, normalization layers and max-pooling layers are also applied. We utilize the fc7 layer (the last fully-connected layer) as the input of UMF-H. The final loss is computed by specifying the relative importance of attribute prediction and categorization. The structure of our model is illustrated in Fig.4. We denote the whole method as UMF-Deep.

During optimization, we first initialize the network parameters with an off-the-shelf model [24] which is pre-trained with ImageNet object recognition dataset. The new added parameters are randomly sampled uniform data based on xavier [50]. Then we fine-tune our model by setting the learning rate on the newly introduced layers larger than the already learned layers. Referring to the original training setting, we use Stochastic Gradient Descent with momentum of to update the weights.

5 Experiments

5.1 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 ImageNet attributes and Caltech-UCSD Birds 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) [11]. 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) [1]. 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) [4]. This dataset contains 11,788 images of 200 bird classes. Each image is annotated with 312 attributes and the category-attribute associations are also available. As a result, this dataset can be utilized for both attribute prediction and zero-shot learning. Since this dataset gives a fine-grained category description, it seems harder to leverage category information to promote instance-level attribute learning than AWA.

ImageNet Attributes (INA) [2]. 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. In addition, λ_1, λ_2, λ_3 are constrained to be equal without tuning separately. We use the 4096-D DeCAF features which are extracted by the Convolutional Neural Networks (CNN) described in [24]. The performance of attribute predictors are measured by *mean area under ROC curve* (mAUC).

For Zero-shot learning, we use the specified seen and

0162-8828 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Copyright (c) 2018 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2836461, IEEE
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2018.2836461
Transactions on Pattern Analysis and Machine Intelligence

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 the first 150 classes as seen classes and leave the remaining 50 classes as unseen. For AwA dataset, we use DeCAF features provided on the dataset website [1] . As for CUB, we extract 4096-D DeCAF features using the method in [24]. We also use the superior VGG19 features [51] for further comparison. 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*.

5.2 Instance-level Attribute Prediction

5.2.1 Attribute Prediction with Known Category Label

We first validate whether category label can benefit attribute prediction. In this setting, the ground truth category label of a test image is provided to do attribute prediction. We choose aPascal-aYahoo and Caltech-UCSD Birds datasets to compare the proposed UMF-CS with some related methods including:

- Direct attribute prediction (DAP). We use linear logistic regression to train attribute classifiers separately. The optimal value for weight decay term is chosen on the validation set.
- Direct concatenation method (Concat.). We simply concatenates the image and category label, $\mathbf{X}$ and $\mathbf{Y}$, as the input for attribute learning. This method is a strong baseline, since it leverages the category information in an additive way.

For the proposed UMF-CS, the optimal value of λ_1, λ_2, λ_3 are 10^{-4} on both aPascal-aYahoo and Caltech-UCSD Birds datasets. As shown in Table 1, the performance of attribute prediction is improved on both datasets by leverage category labels. This proves that category information is an essential factor for learning attributes. Moreover the proposed method achieves better or comparable performance on aPaY and CUB datasets. In addition, the simple concatenation method can perform fairly good when the input information is reliable and sufficient.

We also visualize some results of aPaY dataset in Fig.5 for understanding how category information can enhance attribute prediction. A very common problem of direct attribute prediction is that there may be multiple objects appearing in the same image, e.g. a person sitting on a car, a person riding a horse. As a result, the attribute classifiers do not know which object is referred to be predicted. Another problem is the attribute visual manifestation across categories are large. For example, a monkey's arm is quite different from a statue's arm and a bird head varies from a person's head. Therefore it is more appropriate to tackle attribute prediction in a category sensitive manner.

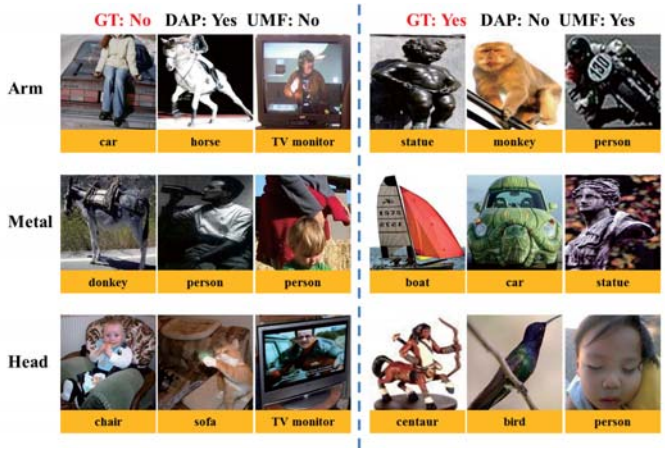


Fig. 5. Corrected attribute prediction by leveraging category information. The yellow box below the image shows the category label of the object.

TABLE 1
Attribute prediction (mAUC) on aPaY and CUB datasets. "Use Categ." indicates whether the method need category labels during training stage. "Know Categ." indicates whether category information is known during test stage. "$*$" indicates the experimental results where the category labels of test images are observed.

Methods	aPaY	CUB	Use/Know Categ.
DAP	0.9367	0.7520	No/No
Concat.	0.9337 0.9745*	0.7615 0.7766*	Yes/No Yes/Yes
MT	0.9395	0.7627	Yes/No
AGL	0.9373	0.7585	Yes/No
HAP	0.9271	0.7282	Yes/No
CSHAP	0.9279	0.7301	Yes/No
UMF-CS	0.9303 0.9742*	0.7659 0.7790*	Yes/No Yes/Yes
UMF-IS	0.9390	0.7672	Yes/No
UMF-H	**0.9413**	**0.7692**	Yes/No

5.2.2 Unifying Attribute Prediction with Categorization

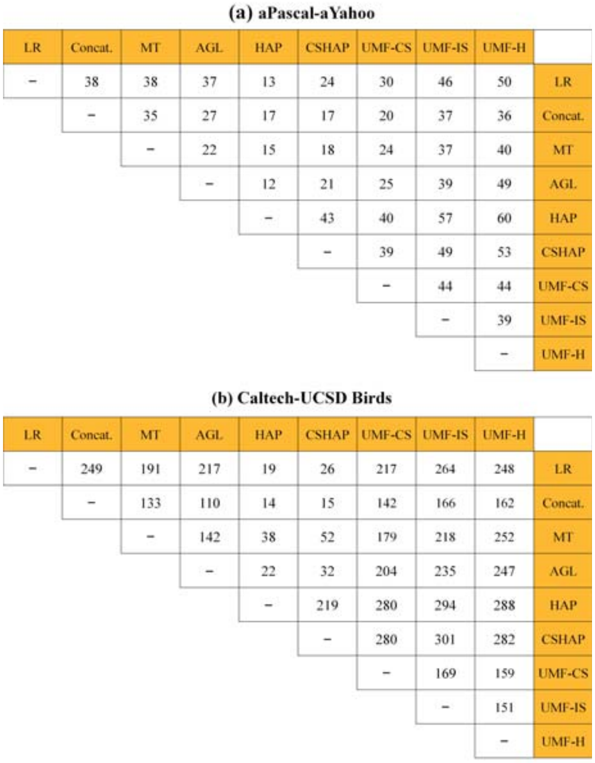
In this experiment, we unified the two tasks: attribute learning and categorization into a joint framework. And we compare the proposed methods (UMF-CS, UMF-IS and UMF-IS-H) with some related methods including: Direct Attribute Prediction (DAP) as introduced in the above experiment, direct Concatenation method (Concat.), Multi-task learning method (MT), Attribute Group Lasso based multi-task learning method (AGL) [13] , Hypergraph-based Attribute Predictor (HAP) [15] and Class-Specific Hypergraph-based Attribute Predictor(CSHAP) [15].

- Direct concatenation method (Concat.). Since the ground truth category label is not provided in this setting, we first train a softmax classifier to do multi-class categorization. Then we concatenate the image feature and category prediction of the softmax classifier, $\mathbf{X}$ and $\tilde{\mathbf{Y}}$, as the input of multiple logistic regression for attribute learning.
- Multi-task learning method (MT). We cascade a hidden layer after the input image feature and the hidden units are further used as the shared input feature of multiple logistic regression and softmax regression fxor attribute learning and multi-class categorization respectively.

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

0162-8828 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Copyright (c) 2018 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2836461, IEEE Transactions on Pattern Analysis and Machine Intelligence
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2018.2836461



(a) aPascal-aYahoo

LR	Concat.	MT	AGL	HAP	CSHAP	UMF-CS	UMF-IS	UMF-H	
–	38	38	37	13	24	30	46	50	LR
	–	35	27	17	17	20	37	36	Concat.
		–	22	15	18	24	37	40	MT
			–	12	21	25	39	49	AGL
				–	43	40	57	60	HAP
					–	39	49	53	CSHAP
						–	44	44	UMF-CS
							–	39	UMF-IS
								–	UMF-H

(b) Caltech-UCSD Birds

LR	Concat.	MT	AGL	HAP	CSHAP	UMF-CS	UMF-IS	UMF-H	
–	249	191	217	19	26	217	264	248	LR
	–	133	110	14	15	142	166	162	Concat.
		–	142	38	52	179	218	252	MT
			–	22	32	204	235	247	AGL
				–	219	280	294	288	HAP
					–	280	301	282	CSHAP
						–	169	159	UMF-CS
							–	151	UMF-IS
								–	UMF-H

Fig. 6. Pairwise comparison of the relevant methods. Each cell counts the number of attributes which are predicted better by the topmost method than the rightmost method. For each attribute, we choose area under ROC curve (AUC) to evaluate the comparison methods.

- Attribute Group Lasso based multi-task learning method (AGL). AGL is a multi-task learning based approach which uses structured sparsity to encourage feature competition among unrelated attributes and feature sharing among related attributes. We use the predefined attribute group [52] as the side information for semantic relatedness.
- Hypergraph-based Attribute Predictor (HAP) and Class-Specific Hypergraph-based Attribute Predictor(CSHAP). HAP proposes to use a hypergraph for depicting the attribute relations in the data which is further casted as a regularized hypergraph cut problem for attribute predication. By considering HAP as a multi-graph cut task, CSHAP can flexibly incorporate category label as side information.

For AGL, HAP and CSHAP, we use the publicly available codes to run the experiments. The hyper-parameters are tuned based on the validation set. The optimal values of $\lambda_i, i = 1, 2, 3$ are 10^{-5}, 10^{-3} and 10^{-4} for UMF-CS, UMF-IS and UMF-H on aPaY dataset. As for CUB, the optimal values of parameter $\lambda_i, i = 1, 2, 3$ are 10^{-4} for both UMF-CS and UMF-IS and 10^{-3} for UMF-H. For λ_4, the optimal values are 10^{-5} and 10^{-6} on aPaY and CUB respectively.

On aPascal-aYahoo and Caltech-UCSD Birds datasets, the proposed method UMF-H achieves the best performance as shown in Table 1. Comparing UMF-IS with UMF-H, we can see that attribute prediction based on the hidden representation of category prediction results in better performance than using the category prediction directly. In addition, the attribute prediction performance of Concat. and UMF-CS decreases on aPascal-aYahoo dataset by considering category information. Therefore, tackling the uncertainty of category prediction is a key factor to improve attribute prediction. Since Caltech-UCSD Birds is a fine-grained dataset, the visual attribute manifestations across categories vary less than aPascal-aYahoo. As a result, the category prediction uncertainty problem is not that serious and Concat. and UMF-CS both achieve better performance than LR. From our perspective, the reason why the proposed method achieves better performance than the multi-task learning method is that the architecture of UMF is asymmetric. More specifically, since our goal is to improve attribute prediction, categorization should only be considered as an auxiliary task. On the contrary, multi-task learning method tackles these two tasks symmetrically. AGL achieves better performance than LR which shows the effectiveness of attribute group information on attribute prediction. Comparing MT and AGL, MT performs slightly better which means the category label is more useful than attribute group information. Comparing HAP and CSHAP, we can find that category-specific attribute predictor can always achieve better performance.

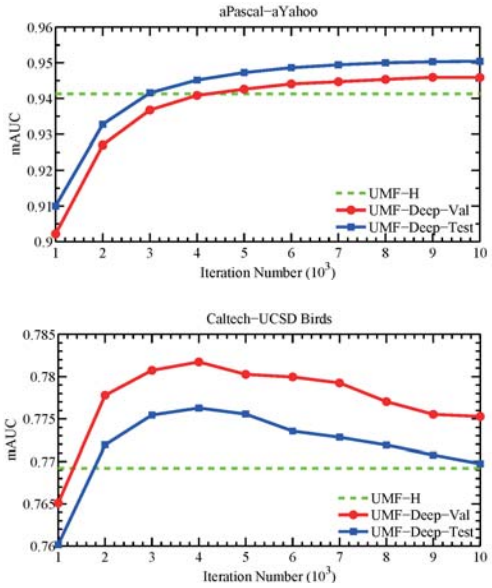


Fig. 7. Deep attribute prediction on aPascal-aYahoo and Caltech-UCSD Birds datasets.

We also give a detail comparison in Fig.6 for all the methods in a pairwise way. Comparing UMF-H and LR, 50 out of 64 and 248 out of 312 attributes are promoted on aPascal-aYahoo and Caltech-UCSD Birds datasets respectively.

5.2.3 *Deep Learning Instance-level Attribute*

By integrating UMF with a deep learning network, we can both accurately predicting attributes and learn the optimal representation of images. We train the UMF-Deep defined in Section 4.3. The base learning rates for SGD solver are set as 10^{-4} for both aPaY and CUB. The learning rate of the new added layers are set 100 times bigger than the layers

0162-8828 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Copyright (c) 2018 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2836461, IEEE Transactions on Pattern Analysis and Machine Intelligence
This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication. The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2018.2836461

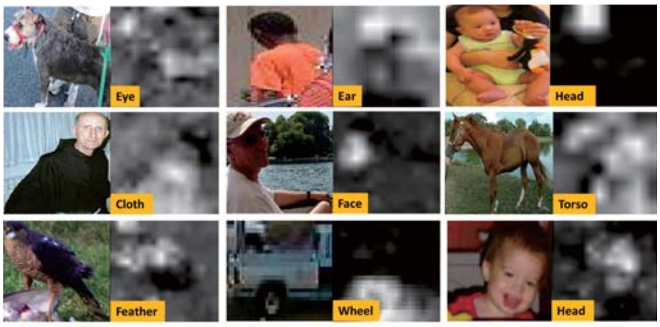


Fig. 8. Attention area of UMF-Deep attribute predictor on aPaY. The yellow box indicates the attribute for visualization.

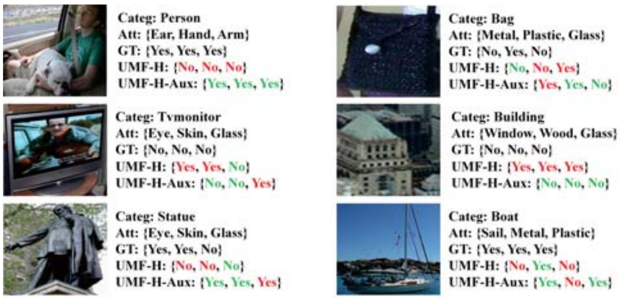


Fig. 9. Enhanced attribute prediction on aPascal-aYahoo dataset by leveraging auxiliary data.

belonging to the off-the-shelf framework. The weight decay term is set as 0.0005 without further tuning. We use the same batch size with the setting of the pretrained model. The other parameters take the same values as those of UMF-H model in the former experiments.

From Fig.7, we can see the optimal value of mAUC is 0.9503 after training 9000 iterations on aPascal-aYahoo dataset. For Caltech-UCSD Birds, the mAUC of UMF-Deep increases first and achieves the optimal value 0.7763 around 3000 iterations. Then the performance decreases because of overfitting problem. Compared with its original version (UMF-H) of the proposed method, UMF-Deep achieves a relatively large increase on both datasets. We also visualize the attention area for the proposed UMF-Deep attribute predictor by masking part of the image [53]. As shown in Fig. 8, most of the attention areas are correlated with the attributes for prediction.

5.2.4 *Enhancing Instance-level Attribute Prediction*

In this experiment, we validate the effectiveness of enhancing attribute prediction ability on aPascal-aYahoo dataset by leveraging auxiliary data which are annotated with only category labels. To this end, we equally separate the origin training data into two parts. The first part is used as the new training data with both attribute labels and category labels. The second part serves as the auxiliary data with only category labels. Then we train the proposed models defined in Eqn. (14) and Eqn. (15) which are named as UMF-IS-Aux and UMF-H-Aux respectively. For UMF-IS and UMF-IS-Aux, λ_*, β are chosen to be 10^{-3} and 0.1. The optimal value of λ_*, β for UMF-H and UMF-H-Aux are 10^{-4} and 0.5.

TABLE 2
Attribute prediction on aPaY dataset with auxiliary data

Methods	mAUC	# Improved Attributes
UMF-IS	0.9313	51 out of 64
UMF-IS-Aux	0.9339	
UMF-H	0.9323	54 out of 64
UMF-H-Aux	0.9389	

The experimental results are shown in Table 2. Among all the 64 attributes in aPascal-aYahoo dataset, the attribute prediction performance of 51 and 54 attributes are improved by UMF-IS-Aux and UMF-H-Aux respectively. Compared with UMF-IS-Aux, UMF-H-Aux achieves better improvement gap by leveraging the same auxiliary data. Then we sample some test images in Fig.9. The attribute predictions of those images are changed relatively large by introducing the auxiliary data. We also show attribute prediction performance of UMF-H and UMF-H-Aux for individual attribute in Fig.10. Attributes named ”Leather”, ”Round”, ”Wool”, ”Sail”, ”Occluded”, ”2D Boxy” are enhanced most with the auxiliary information. From the result we can see that some attributes are predicted better due to the more reliable category information. Since the annotation of instance-level attribute is costly, enhancing attribute prediction by using auxiliary task such as object recognition seems promising.

5.3 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 Fig. 11. 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

TABLE 3
Attribute prediction on category-attribute pairs.

Methods	aPay		CUB	
	mAUC	# Classifiers	mAUC	# Classifiers
U	0.584	64	0.770	25
CS	0.618	383	0.843	607
UMF-CS	0.600	96	0.841	409
UMF-IS	0.603	96	0.847	409

0162-8828 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Copyright (c) 2018 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2836461, IEEE Transactions on Pattern Analysis and Machine Intelligence
This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2018.2836461

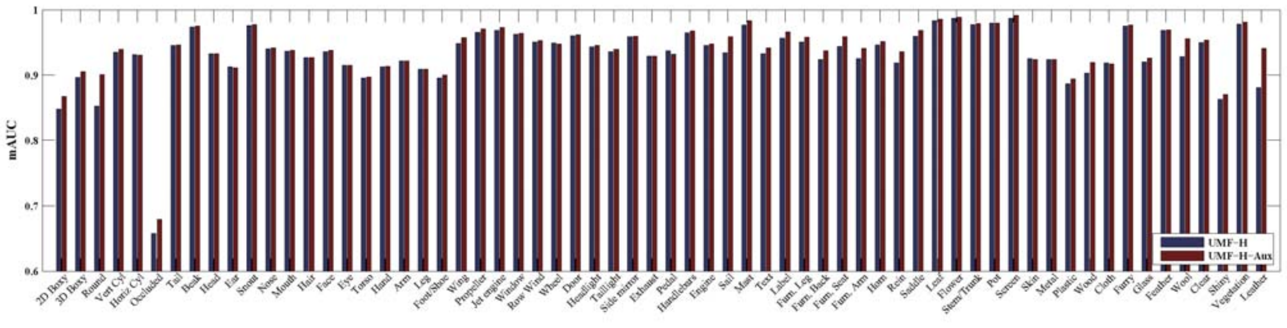


Fig. 10. Quality of individual attribute predictors trained with UMF-H by leveraging auxiliary data.

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 attributes in a category-dependent way. To train the t^{th} attribute classifier for category j, the violating constraints for positive and negative samples from category j are given a higher penalty, as suggested in [34].

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.

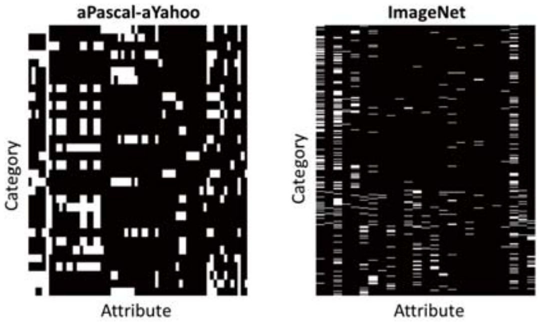


Fig. 11. Valid category-attribute pairs. The white entities denote the category-attribtue pairs have both positive and negative exemplars.

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 attributes 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.

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

Methods	AWA	CUB	Fea Type
DAP	41.4 [11]/45.3*	−/16.9*	H/D
IAP	42.2 [11]/46.4*	−/16.7*	H/D
ALE	37.4 [27]/45.7 [26]	18.0 [27]/20.2 [26]	FV/D
BN	43.4 [14]	−	H
TMV-HLP	47.1 [29]	−	H
HAP-H	45.0 [15]	17.5 [15]	D
HAP-G	45.0 [15]	17.5 [15]	D
UDCIA	63.6 [35]	42.4 [35]	V
KDCIA	73.8 [35]	43.7 [35]	V
UMF-IS w/o $\mathbf{W}$	43.7	15.9	D
UMF-CS	47.1	17.7	D
UMF-IS	48.6/62.7	18.2/42.0	D/V

5.4 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* [11]. 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:

$$P(z_l|\mathbf{x}) = \frac{p(z_l)}{p(a^{z_l})} \prod_{t=1}^{T} p(a_t^{z_l}|\mathbf{x}), \tag{16}$$

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

0162-8828 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Copyright (c) 2018 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2836461, IEEE Transactions on Pattern Analysis and Machine Intelligence

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication. The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2018.2836461

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:

$$p(a_t^{z_l}|\mathbf{x}, \tilde{\mathbf{y}}) = \frac{[\![a_t^{z_l} = 1]\!]}{1 + e^{-f(\mathbf{x},\tilde{\mathbf{y}},t)}} + \frac{[\![a_t^{z_l} = 0]\!]}{1 + e^{f(\mathbf{x},\tilde{\mathbf{y}},t)}} \tag{17}$$

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) [14], transductive multi-view Bayesian label propagation (TMV-BLP) [29], Hypergraph-regularized Attribute Predictors (HAP) [15], UDCIA [35] and KDCIA [35]. 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, by using the same DeCAF feature, 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. With the superior VGG19 feature, the performance of our proposed method can be further improved.

5.5 Recognizing Human-Object Interactions

Beyond attribute learning, we also apply the proposed method to recognize human object interactions (HOI). For an input still image, the algorithm is expected to output the category of human actions (e.g. "riding a bike", "boarding an airplane"). HOI recognition is highly correlated to typical attribute learning. The difference lies in that actions are predicted for each object as its "dynamic" attributes in HOI recognition, instead of "still" attributes in typical attribute learning. These two tasks are the prerequisite for high-level image semantic understanding such as image caption generation and visual question answering.

We evaluate the proposed method on the recently released HICO dataset [54]. This dataset contains 47774 images and 600 human object interactions. Each image is annotated with 80 object categories and 117 actions. Different from the datasets for attribute learning, multiple objects may be contained in a single image. So we replace the softmax classifier with multiple logistic regression for our method to leverage category information. We sum up the scores of object recognition classifier and action recognition classifier to judge how human interact with the object.

Several approaches are presented in the HICO benchmark: Fisher Vectors [55], Random Forest [56], DNN features from AlexNet [17] and HOCNN (Human-Object C-NN) [54]. Except for using the off-the-shelf DNN features from AlexNet, there are three variants by using different fine-tuning strategies. HOCNN uses the outputs of object detection [18] trained on MS-COCO [57] and human pose estimation [58] as features. Following the original paper, we use mean average precision (mAP) as the evaluation metric.

We compare UMF-deep with the above methods. The number of hidden units is set to 500 and the other paramters are the same with Section 5.2.3. The data with ambiguous label are filtered out for evaluation at test stage. From the experimental results in Table 5, we can see that the feature learned by deep convolutional neural network surpasses the other shallow feature with a large gap. Among all the three fine-tuning strategies, fine-tuning the network with object labels achieves better performance than using the other two information (e.g. action and object-action pair). This shows that object information is an essential factor in recognizing human object interactions. By elaborately modelling the relationship between objects and actions, our method achieves the best performance among all the comparison methods. This further verifies the importance of object recognition in HOI task.

TABLE 5
Performances on HICO dataset.

Methods	mAP	Fea Type
Random	0.57	–
Fisher Vectors	4.21	shallow
Random Forest	7.30	shallow
HOCNN	4.90	deep
DNN(ImageNet)	18.58	deep
DNN(fine-tune V)	17.65	deep
DNN(fine-tune VO)	18.08	deep
DNN(fine-tune O)	19.38	deep
UMF-Deep	**20.45**	deep

6 Conclusions

We propose a unified multiplicative framework for attribute learning by leveraging category information. Comparing to 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.

Acknowledgements

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

References

[1] 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*. IEEE, 2009, pp. 1778–1785.

[2] 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.

0162-8828 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Copyright (c) 2018 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2836461, IEEE Transactions on Pattern Analysis and Machine Intelligence

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication. The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2018.2836461

[3] W. Ouyang, H. Li, X. Zeng, and X. Wang, “Learning deep representation with large-scale attributes,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1895–1903.

[4] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The Caltech-UCSD Birds-200-2011 Dataset,” California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[5] 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*. IEEE, 2009, pp. 365–372.

[6] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.

[7] 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*. IEEE, 2012, pp. 2973–2980.

[8] Y. Wang and G. Mori, “A discriminative latent model of object classes and attributes,” in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 155–168.

[9] D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith, “Default probability,” *Cognitive Science*, vol. 15, no. 2, pp. 251–269, 1991.

[10] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, “Learning systems of concepts with an infinite relational model,” in *AAAI*, vol. 3, 2006, p. 5.

[11] 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*, vol. 36, no. 3, pp. 453–465, 2014.

[12] 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*. IEEE, 2014, pp. 1027–1034.

[13] 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*. IEEE, 2014, pp. 1629–1636.

[14] X. Wang and Q. Ji, “A unified probabilistic approach modeling relationships between attributes and objects,” in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2120–2127.

[15] S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang, “Learning hypergraph-regularized attribute predictors,” in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2015.

[16] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, “Scheduled sampling for sequence prediction with recurrent neural networks,” in *Advances in Neural Information Processing Systems*, 2015, pp. 1171–1179.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *arXiv preprint arXiv:1512.03385*, 2015.

[20] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[21] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, “Part-based r-cnns for fine-grained category detection,” in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 834–849.

[22] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[23] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “Cnn features off-the-shelf: an astounding baseline for recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.

[24] 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*, 2014, pp. 647–655.

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

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

[27] 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*. IEEE, 2013, pp. 819–826.

[28] S. J. Hwang and L. Sigal, “A unified semantic embedding: Relating taxonomies and attributes,” in *Advances in Neural Information Processing Systems*, 2014, pp. 271–279.

[29] 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*. Springer, 2014, pp. 584–599.

[30] X. Wang, J. Bi, S. Yu, and J. Sun, “On multiplicative multitask feature learning,” in *Advances in Neural Information Processing Systems*, 2014, pp. 2411–2419.

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

[32] 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*, 2014, pp. 2348–2356.

[33] S. J. Hwang, S. Fei, and K. Grauman, “Sharing features between objects and their attributes.” in *Computer Vision and Pattern Recognition*, 2011, pp. 1761–1768.

[34] C.-Y. Chen and K. Grauman, “Inferring analogous attributes,” in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 200–207.

[35] C. Gan, T. Yang, and B. Gong, “Learning attributes equals multi-source domain generalization,” in *Computer Vision and Pattern Recognition*, 2016, pp. 87–97.

[36] J. Lei Ba, K. Swersky, S. Fidler *et al.*, “Predicting deep zero-shot convolutional neural networks using textual descriptions,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4247–4255.

[37] Y. Fu, T. Hospedales, T. Xiang, and S. Gong, “Transductive multi-view zero-shot learning,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[38] D. Jayaraman and K. Grauman, “Zero-shot recognition with unreliable attributes,” in *Advances in Neural Information Processing Systems*, 2014, pp. 3464–3472.

[39] S. Shankar, V. K. Garg, and R. Cipolla, “Deep-carving: Discovering visual attributes by carving deep neural nets,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3403–3412.

[40] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, “Panda: Pose aligned networks for deep attribute modeling,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1637–1644.

[41] W. T. Freeman and J. B. Tenenbaum, “Learning bilinear models for two-factor problems in vision,” in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*. IEEE, 1997, pp. 554–560.

[42] C. G. Snoek, M. Worring, and A. W. Smeulders, “Early versus late fusion in semantic video analysis,” in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 399–402.

[43] Y. Dong, S. Gao, K. Tao, J. Liu, and H. Wang, “Performance evaluation of early and late fusion methods for generic semantics indexing,” *Pattern Analysis and Applications*, vol. 17, no. 1, pp. 37–50, 2014.

[44] D. Liu, K.-T. Lai, G. Ye, M.-S. Chen, and S.-F. Chang, “Sample-specific late fusion for visual category recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 803–810.

[45] O. R. Terrades, E. Valveny, and S. Tabbone, “Optimal classifier fusion in a non-bayesian probabilistic framework,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 9, pp. 1630–1644, 2009.

[46] G. Ye, D. Liu, I.-H. Jhuo, S.-F. Chang *et al.*, “Robust late fusion with rank minimization,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3021–3028.

0162-8828 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Copyright (c) 2018 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2836461, IEEE
Transactions on Pattern Analysis and Machine Intelligence
This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TPAMI.2018.2836461

[47] K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain, "Likelihood ratio-based biometric score fusion," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 2, pp. 342–347, 2008.

[48] J. Liu, S. McCloskey, and Y. Liu, "Local expert forest of score fusion for video event classification," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 397–410.

[49] 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*. Springer, 2008, pp. 117–132.

[50] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International conference on artificial intelligence and statistics*, 2010, pp. 249–256.

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

[52] C. Lampert, "Semantic attributes for object categorization (slides)," in *http://ist.ac.at/ chl/talks/lampert-vrml2011b.pdf*, 2011.

[53] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 740–755.

[54] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "Hico: A benchmark for recognizing human-object interactions in images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1017–1025.

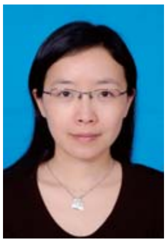
[55] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.

[56] B. Yao, A. Khosla, and L. Fei-Fei, "Combining randomization and discrimination for fine-grained image categorization," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1577–1584.

[57] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 740–755.

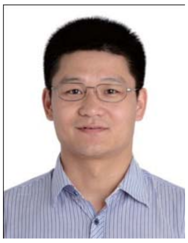
[58] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Advances in Neural Information Processing Systems*, 2014, pp. 1736–1744.

Kongming Liang received the Bachelor's degree from China University of Mining & Technology-Beijing, China, in 2012; Currently, he is a Ph.D candidate in Institute of Computing Technology, Chinese Academy of Science since 2012. From Sep 2016 to Oct 2017, he was a joint Ph.D. Student of machine learning group in Carleton University, Canada. His research interests cover computer vision and machine learning, especially visual attribute learning, visual relationship detection and holistic image understanding based on deep neural networks.



Hong Chang received the Bachelor's degree from Hebei University of Technology, Tianjin, China, in 1998; the M.S. degree from Tianjin University, Tianjin, in 2001; and the Ph.D. degree from Hong Kong University of Science and Technology, Kowloon, Hong Kong, in 2006, all in computer science. She was a Research Scientist with Xerox Research Centre Europe. She is currently an Associate Researcher with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. Her main research interests include algorithms and models in machine learning, and their applications in pattern recognition, computer vision, and data mining.

Bingpeng Ma received the B.S. degree in mechanics in 1998 and the M.S. degree in mathematics in 2003 from Huazhong University of Science and Technology, respectively. He received Ph.D. degree in computer science at the Institute of Computing Technology, Chinese Academy of Sciences, P.R. China in 2009. He was a post-doctorial researcher in University of Caen, France, from 2011 to 2012. He joined the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, in March 2013 and now he is an associate professor. His research interests cover computer vision, pattern recognition, and machine learning. He especially focuses on face recognition, person re-identification and the related research topics.



Shiguang Shan received M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, and Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2004. He joined ICT, CAS in 2002 and has been a Professor since 2010. He is now the deputy director of the Key Lab of Intelligent Information Processing of CAS. His research interests cover computer vision, pattern recognition, and machine learning. He especially focuses on face recognition related research topics. He has published more than 200 papers in refereed journals and proceedings in the areas of computer vision and pattern recognition. He has served as Area Chair for many international conferences including ICCV'11, ICPR'12, ACCV'12, FG'13, ICPR'14, ICASSP'14, and ACCV'16. He is Associate Editors of several international journals including IEEE Trans. on Image Processing, Computer Vision and Image Understanding, Neurocomputing, and Pattern Recognition Letters. He is a recipient of the China's State Natural Science Award in 2015, and the China's State S&T Progress Award in 2005 for his research work.

Xilin Chen received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1988, 1991, and 1994, respectively. He was a professor with the Harbin Institute of Technology from 1999 to 2005. He has been a professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS), since August 2004. He is the Director of the Key Laboratory of Intelligent Information Processing, CAS. He has published one book and over 200 papers in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. He is now an associate editor of IEEE Transactions on Multimedia, Journal of Visual Communication and Image Representation, a leading editor of Journal of Computer Science of Technology, and the associate editor in chief of the Chinese Journal of Computer. He is a recipient of several awards, including the Chinas State Natural Science Award in 2015, the Chinas State Scientific and Technological Progress Award in 2000, 2003, 2005, and 2012 for his research work. He is a Fellow of IEEE / IAPR / CCF.

0162-8828 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Copyright (c) 2018 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.