Clustering Multi-task Feature Learning for Attribute Prediction

Anonymous CVPR submission

Paper ID 1105

Abstract

Semantic attributes have been proposed to bridge the semantic gap between low-level feature representation and high-level semantic understanding of visual objects. Obtaining a good representation of semantic attributes usually requires learning from high-dimensional low-level features, which often suffers from the curse of dimensionality. Designing a good feature-selection approach would benefit attribute prediction and in turn its related applications. Since semantic attributes of an object are usually “related”, in the literature multi-task learning has been introduced for multi-attribute prediction, either by assuming that all attributes are somehow correlated or by manually dividing attributes into related groups. However, the performance of such approaches greatly rely on the task structure. The prediction performance would degrade if the assumed task structure does not match to that of the problem. Desired is an approach that can automatically detect problem-specific clustering structures of the attributes. In this paper, we propose a novel clustered multi-task feature selection approach utilizing K-means and group sparsity regularizers, and develop an efficient alternating optimization algorithm. Experiments demonstrate that the proposed approach can automatically capture the task structure and hence result in obvious performance gain in attribute prediction, when compared with existing state-of-the-art approaches.

1. Introduction

Recent literature has witnessed fast development of representations using semantic attributes, whose goal is to bridge the semantic gap between low-level feature representation and high-level semantic understanding of visual objects. Attributes refer to visual properties that help describe visual objects or scenes such as “natural” scenes, “fluffy” dogs, or “formal” shoes. Visual attributes exist across object category boundaries and many methods have been employed in applications including object recognition [7, 6], face verification [23] and image search [17, 21].

In a real-world problem, one attribute is often related to some other attributes. For example, as shown in Figure 1, a “high-heel” shoe is usually considered as a “formal” shoe as well. To generate better generalization performance by capturing such correlation, recent work [3, 13] started introducing multi-task learning approaches into attribute prediction or ranking. However, if one assumes such correlation exists across all attributes, the assumption would be too strong. For example, it is hard to identify whether “high heel” or “formal” shoes are in red. This suggests that in real applications the correlation may only exist within sub-groups of the attributes. A naive approach of learning attributes from all the groups jointly, no matter whether they are truly related, would obviously lead to less-than-optimal performance due to the unnecessary and incorrect constraints. On the other hand, manually defining grouping/clustering structures for the attributes may be possible only for very simple problems with a small number of attributes. In short, existing approaches still lack the capability of automatically detecting grouping structures of the attributes for facilitating learning.

Good representations of semantic attributes are often built on top of high-dimensional, low-level features. Attribute learning directly based on such raw, high-dimensional features may suffer from the problem of dimensionality curse. Further, often it is reasonable to assume that not all the low-level features would have equal contribution to all the attributes. In other words, there may be clustering structures existing across the dimensions as well. Identifying grouping/clustering structures both among the attributes and across the dimensions would naturally contribute to improved learning. The latter is essentially an attribute-dependent feature selection problem. In this paper,
we propose a regularization-based multi-task-learning approach that aims at automatically partitioning the attributes into groups while simultaneously utilizing such group structures for attribute-dependent feature selection. We employ a K-means regularizer for attribute clustering, where strong attribute correlation is assumed to exist within each cluster. Besides, a group-sparsity regularizer is imposed on the objective function to encourage intra-cluster feature sharing and inter-cluster feature competition. Under this formulation, we also propose an alternating structure optimization algorithm, which efficiently solves the proposed formulation. We verify the effectiveness and generalization capability of our approach on one synthetic dataset and three real image datasets. The results show that our approach outperforms the state-of-the-art approaches on prediction accuracy and zero-shot learning.

In the remaining of the paper, we first discuss related work in Section 2. The proposed approach is presented in Section 3. Experiments and results are demonstrated in Section 4. We conclude the paper in Section 5.

**Notations:** In this paper, we represent scalars, vectors, matrices and sets as lower case letters $x$, bold face lower case letters $x$, capital letters $X$ and calligraphic capital letters $\mathcal{X}$ respectively. $x_i$ and $x_{(ij)}$ denote the $i$-th column and the $j$-th row of the matrix $X$. The scalar $x_{ij}$ denotes the $(i, j)$-th entry of $X$. $tr(X)$ denotes the trace of $X$. $\| \cdot \|_p$ and $\| \cdot \|_F$ represent $p$-norm and Frobenius norm respectively.

### 2. Related Work

As our work is mostly related to multi-task learning and attribute learning, we briefly review the literature on these two topics in the following and draw the conclusion that neither current multi-task learning nor attribute learning approaches can be adopted directly to automatically partition the attributes into different groups based on correlations for feature selection.

#### 2.1. Multi-task Learning

Assuming several different but similar tasks are “related”, multi-task learning aims to learn several tasks together to improve the generalization performance through capturing the intrinsic correlation. Multi-task learning has been successfully applied to many computer vision applications including visual classification [26], action recognition [19], attribute learning [3, 13], etc.

There have been two main ways to define task relatedness. The first one assumes that all tasks share a similar parameter space. Ji and Ye [14] introduced trace norm as regularization and obtained a low-rank structure projection matrix to capture task relatedness. Bach [2] and Jacob et al. [11] assume the tasks have some special structures and assign tasks into different groups where tasks in the same group are closer to each other than tasks in a different group. Kanget al. [15] assign tasks into groups through integer programming. Kim and Xing [16] manage tasks in a tree structure where tasks from the same node are closer to each other and relatedness among the nodes depends on the depth in a tree. Similarly, Chen et al. [4] represents tasks in a graph where task relatedness depends on the edge weight between two nodes. Some of these approaches considered structures of tasks, but they serve only to learn the shared parameter space.

The other way models task relatedness as a common subset of latent features shared by different tasks. Argyriou et al. [1] obtain a sparse projection matrix through a $\ell_1/\ell_q$ norm group lasso regularizer. Jacob et al. [10] further introduce a graph of covariates as prior for feature selection. Zhang et al. [28] provides a probabilistic interpretation of an appropriate q in the generalized $\ell_1/\ell_q$ norm. Jalali et al. [12] and Gong et al. [9] introduce an extra $\ell_1$ and $\ell_1/\ell_q$-norm regularization term individually to detect outliers. However, current multi-task feature learning approaches select features by treating all tasks in a whole group based on the assumption that all tasks are strongly correlated, thus cannot be directly adopted in our problem.

#### 2.2. Attribute Learning

A visual attribute learner is a binary predictor that aims to indicate whether or not a visual property is present. The standard approaches learn the attribute predictor independently per attribute. Ferrari and Zisserman [8] presented a probabilistic generative model which learns attributes by distinguishing unary property of single segment or patterns of alternating segments. Lampert et al. [18] considered zero-shot learning where the test set consists of entirely previously unseen object categories and the information is transferred from the training set to the test phase entirely through the attribute labels. Farhadi et al. [7] described unfamiliar objects and new categories by visual attribute of object parts, e.g., “has head”, or appearance adjectives, e.g., “spotty”. Farhadi et al. [6] first learned part and category detectors of objects and then described objects by spatial arrangement of the attributes and their interactions. Kovashka et al. [17] and Scheirer et al. [21] used attributes to facilitate human-machine interaction for image search by which the user is able to specify precise semantic queries.

While most methods learn attributes independently, some initial steps have been taken towards modeling attribute relationships. Wang et al. [25] treated attributes as latent variables and capture the correlations among attributes using an undirected graphical model built from training data. Song et al. [23] proposed a method to model the attribute relationship for face verification based on a discriminative distributed-representation for attribute description. Siddiquie et al. [22] proposed retrieval approach where correlations of attributes are considered as multi-
attributes query in the vocabulary.

Considering utilizing the multi-task learning framework for attribute learning, Chen et al. [3] proposed a ranking framework which learns a common feature space among all attributes while detecting outliers; Jayaraman et al. [13] aimed to select appropriate subset of features for different attributes by manually dividing attributes into different semantic groups and encouraging intra-group feature-sharing and inter-group feature competition.

These approaches either make the strong assumption that all tasks are correlated or require human intervention to specify appropriate semantic groups. In contrast, our approach can automatically detect semantic groups and learns an effective subset of features representing the attributes.

3. Methodology

In this section, we first give a formal definition of the problem, then we propose a clustered multi-task feature selection framework together with an optimization algorithm to solve the problem.

3.1. Problem Definition

Suppose that we are given a multi-task learning problem with \( m \) tasks (attributes); each task \( i \) is associated with a set of training data of \( d \) dimension and \( n \) samples: \((x_1^i, y_1^i), \ldots, (x_n^i, y_n^i) \subset \mathbb{R}^d \times \mathbb{R}\), we denote \( W = [w_1, \ldots, w_m] \in \mathbb{R}^{d \times m} \) as the projection weight matrix to be estimated where each column \( w_i \) is the weight vector of the \( i \)-th task. Tasks may exhibit grouping structures, as illustrated in Figure 2. The tasks in the same group are highly correlated and thus sharing the same subset dimensions of features. The tasks in different groups are weakly correlated and have different subset of non-zero dimensions of features. Our goal is to design an approach which can automatically detect such group structure and utilize such group correlation information for feature selection.

This problem can be formulated as below:

\[
W, I : \min_{W, I} \mathcal{L}(W | X, Y) + \alpha \mathcal{F}(I) + \beta \mathcal{G}(W)
\]

where \( \mathcal{L}(W | X, Y) \) is the logistic regression objective \( \sum_m \sum_n \log(1 + \exp(-y^i_n (w^T_n \cdot w^m_n))) \), \( I \) is the clustering assignment of the tasks, \( \mathcal{F}() \) is the regularization term to encourage good clustering and \( \mathcal{G}() \) is the regularizer to encourage feature sharing within clusters and feature competition across different clusters.

3.2. Clustered Multi-task Feature Selection

In this subsection, we introduce the proposed choices for the two regularizers. One of the popular regularizers encouraging good clustering is the sum of squared error (SSE) that is used in K-means clustering:

\[
\mathcal{F}(I) = \sum_{j=1}^{k} \sum_{v \in I_j} \|w_v - \bar{w}_j\|_2^2
\]

where \( I_j \) denotes the \( j \)-th cluster whose mean is \( \bar{w}_j \). According to [5, 27], it can also be written as:

\[
\sum_{j=1}^{k} \sum_{v \in I_j} \|w_v - \bar{w}_j\|_2^2 = tr(W^T W) - tr(F^T W^T F)
\]

where the matrix \( F \in \mathbb{R}^{m \times k} \) is an orthogonal cluster indicator matrix with \( F_{i,j} = \frac{1}{\sqrt{n_j}} \) if \( i \in I_j \) and \( F_{i,j} = 0 \) otherwise. This function imposes the constraint that each vector in the same cluster should be close to the mean vector.

By employing the above SSE function and adding an additional term \( tr(W^T W) \) to improve the generalization, [29] derived a relaxed clustered multi-task learning penalty:

\[
M : \min_{M} \alpha \eta (1 + \eta) tr(W(\eta I + M)^{-1} W^T)
\]

s.t. \( tr(M) = k, M \succeq I, M \in \mathbb{S}_+^m \)

where \( M = F^T F \) potentially embeds the cluster assignment information. \( \alpha \) and \( \eta \) are super parameters.

Given the clustering assignment \( I \), the following group regularization is proposed to encourage intra-group feature-sharing and inter-group feature competition:

\[
\mathcal{G}(W) = \sum_{i=1}^d \sum_{j=1}^k \|w_{i(j)}^T\|_2 = \sum_{j=1}^k tr(\sqrt{W_{I_j}W^T})
\]

where \( w_{i(j)}^T \) is a row vector including elements \( w_{ik} \) if \( w_{ik} \in I_j \), \( I_j \) is a diagonal matrix with the element \( I_{ii} = 1 \) if \( w_{ik} \in I_j \) and \( I_{ii} = 0 \) otherwise. This regularizer first imposes an \( \ell_2 \) norm to the row vector of each group to “collapse” each group as a column vector, and then imposes an \( \ell_1 \) norm for sparsity to select features.
Putting all those term together, the proposed objective function for clustered multi-task feature selection with group detection and feature selection is written below:

\[
W, M : \min_{W, M} \mathcal{L}(W|X, Y) + \beta \sum_{j=1}^{k} \text{tr}
\left(\sqrt{W \hat{J}_j W^T}\right) + o(n) (1 + \eta) \text{tr}(W(\eta I + M)^{-1}W^T) \]

s.t. \(\text{tr}(M) = k, M \succeq I, M \in S^n_+\)

where the group information \(\mathcal{I}\) is embedded in \(M\), which will be estimated by the method shown in the subsection below. The key idea lying here is that we use the K-means regularizer to partition the tasks into groups where strong correlation exists among tasks in the same group; and feature selection based on such group structures would make sure appropriate feature subsets are selected to represent the respective semantic attributes.

The method is related to some existing methods in the literature, e.g., [13], which requires the clustering assignment of the attributes \(\mathcal{I}\) as the input. However, the optimal clustering assignment is not always available, if not impossible. A non-optimal input of such clustering assignment will dramatically degrade the performance of the method in [13], which is shown in the experiment section. Instead the proposed method does not require such input, which learns the grouping of the attributes from the tasks automatically, and the learned grouping of the attributes is then used to update the tasks.

### 3.3. Cluster Assignment Identification

Since the K-means regularizer is spectral relaxed, we cannot obtain the cluster assignment information \(\mathcal{I}\) directly from \(M\). In this section we design a procedure to estimate the cluster assignment information, which is summarized in Algorithm 1.

We first need to obtain a good approximation of the cluster indicator matrix \(F\). Given \(M\), we apply Eigen decomposition \(M = U \Lambda U^T\) where each column of \(U\) is the eigenvector and each diagonal element of \(\Lambda\) is the eigenvalue. Then \(n\) columns of \(U\) which have the \(n\) largest corresponding eigenvalues in \(\Lambda\) give an approximation of the cluster assignment matrix \(F_{m \times k}\). The number of cluster can be automatically detected by the absolute value of the eigenvalue. In our approach, we keep all eigenvalues greater than \(10e^{-8}\). Note that the cluster number can also be specified by the user as supervision.

After obtaining an approximation of \(F\), we use the similar technology in [27] to get the assignment information. Specifically, QR decomposition with column pivoting is first applied to \(F\):

\[
F^T = Q[R_{11}, R_{12}]P^T
\]

### Algorithm 1 Obtain cluster assignment information

**Input:** \(M\);

**Output:** Cluster assignment vector \(c\);

1. Apply Eigen decomposition to \(M = U \Lambda U^T\);
2. Obtain \(F\) from \(n\) columns of \(U\) having \(n\) largest eigenvalues;
3. Apply QR decomposition with column pivoting: \(F^T = Q[R_{11}, R_{12}]P^T\);
4. Calculate \(\hat{R} = [I_n, R_{11}^{-1} R_{12}]P^T\);
5. For each task \(i\), \(e_i = \arg \max_j \hat{R}_{ij}\);

where \(Q\) is an \(n \times n\) orthogonal matrix, \(R_{11}\) is an \(n \times n\) upper triangular matrix and \(P\) is a permutation matrix. Then we calculate matrix \(\hat{R}\) by

\[
\hat{R} = [I_n, R_{11}^{-1} R_{12}]P^T.
\]

The assignment information is implied by \(\hat{R}\) where the cluster membership of each task (column) is determined by the row index of the largest element in absolute value of the corresponding column of \(\hat{R}\) (Algorithm 1).

### 3.4. Optimization Algorithm

We propose an alternating optimization algorithm to solve the problem in Eqn. 6, which updates the weight vectors of the tasks \(W\) and the grouping matrix \(M\) alternately until a convergence criterion is satisfied. The whole algorithm is also summarized in Algorithm 2. The details of solving \(W\) and \(M\) are presented in the following.

#### Optimization of \(M\)  

The optimal \(M\) can be obtained via solving:

\[
M : \min_{M} \text{tr}[W(\eta I + M)^{-1}W^T] \quad (9)
\]

s.t. \(\text{tr}(M) = k, M \succeq I, M \in S^n_+\)

Let \(W = U^T V\) be the SVD of \(W\), and \(M = V \Lambda V^T\) be the Eigen decomposition of \(M\), where \(\Lambda\) is a diagonal matrix, then we have

\[
M : \min_{\Lambda} \text{tr}(U \Sigma V^T (\eta I + \Lambda)^{-1} V^T V \Sigma U^T) \quad (10)
\]

s.t. \(\text{tr}(V \Lambda V^T) = k, 0 \leq \Lambda \leq 1\)

This problem is equivalent to the following problem:

\[
\lambda_1, \lambda_2, \ldots, \lambda_q : \min_{\lambda_i} \sum_{i=1}^{q} \frac{\sigma_i^2}{\eta + \lambda_i} \quad (11)
\]

s.t. \(\sum_{i=1}^{q} \lambda_i = k, 0 \leq \lambda_i \leq 1\)

where \(\Lambda = \text{diag}([\lambda_1, \lambda_2, \ldots, \lambda_q])\) and \(\Sigma = \text{diag}([\sigma_1, \sigma_2, \ldots, \sigma_m])\).
Optimization of $W$  By first squaring the mixed-norm regularizer, the objective function can be approximated using the following upper bound:

$$\sum_{i=1}^{d} \sum_{j=1}^{k} \left( \| w_{ij} \|_2 \right)^2 \leq \sum_{i=1}^{d} \sum_{j=1}^{k} \frac{\| w_{ij} \|_2^2}{\delta_{ij}}$$  (12)

where $\delta_{ij}$ are positive dummy variables satisfying $\sum_{i,j} \delta_{ij} = 1$. Then $\delta_{ij}$ can be updated by holding the equality:

$$\delta_{ij} = \frac{\| w_{ij} \|_2}{\sum_{i,j} \| w_{ij} \|_2}$$  (13)

Then for fixed $M$ and $\delta_{ij}$, each weight vector $w$ can be updated by gradient-type approaches by taking the derivative of the following function:

$$\begin{align*}
\mathbf{w} : & \min \mathbb{E}(W|X, Y) + \gamma tr(W(\eta I + M)^{-1}W^T) \\
& + \beta \sum_{i=1}^{d} \sum_{j=1}^{k} \frac{\left( \| w_{ij} \|_2 \right)^2}{\delta_{ij}}
\end{align*}$$  (14)

Optional Imposing cluster structure for supervision
Before calculating $\delta$ after obtaining cluster assignment from $M$, an optional step can be taken if some specific cluster structure is preferred based on the domain knowledge. For example, if by prior we know $w_i$ and $w_j$ should be in the same cluster, an additional operation could be taken on the cluster assignment vector $c$ to ensure those two weight vectors are in the same cluster. Note that not only the whole group structure but also partial group information can be imposed as supervision.

3.5 Complexity Analysis
Following the previous discussion, we use $n$, $d$, $m$ and $k$ to denote the number of instances, the dimensions of the feature, the number of attributes and the number of the clusters. During cluster assignment identification, the computation cost of calculating $M$ includes the eigen decomposition taking $O(dm^2)$, the QR decomposition taking $O(dk^2)$ and the computation of $R$ taking $O(k^3)$. Since $d > m > k$, the total complexity of calculating $M$ by Algorithm 1 takes $O(dm^2)$. For the optimization process, the update of mix-norm takes $O(dk)$. The update of $M$ takes $O(mnd^2) + O(md^2) + O(mnd) + O(dm^2)$. Since $n > d > m$, the total computation complexity in each iteration is $O(mnd^2)$.

The convergence of our algorithm can be proved by using similar techniques from [1].

4. Experiments
In this section, we first verify the effectiveness of our proposed approach in obtaining the correct cluster structures on one synthetic dataset, then we evaluate the attribute prediction and zero-shot learning capability on three real world image datasets.

4.1 Simulation Experiment
The synthetic data set is constructed in a procedure similar to [11, 29]. Specifically, the dataset consists of 5 clusters, where each cluster contains 10 tasks and each task is represented by a weight vector with dimension $d = 30$. Denote by $\hat{w}_c^i$ the weight vector of the $i$-th task from the $c$-th cluster, then $\hat{w}_c^i$ can be expressed as the sum of the cluster center $w^c$ and the task specific component $w^c_i$: $\hat{w}_c^i = w^c + w^c_i$. We then independently generate the cluster center $w^c$ and the task-specific component $w^c_i$.

The cluster centers $w^c$ are first drawn from a normal distribution. Then we randomly set half of the dimensions to zeros. Note that we keep $w^c$ orthogonal to the other cluster centers by selecting the appropriate locations of the non-zero entries. The task-specific component $w^c_i$ is first drawn from the same normal distribution, then the same dimension of feature to their corresponding cluster centers are set to zero.

For each task, we generate 60 samples for training and 1000 samples for testing. Denote the data matrix and the corresponding response as $X_i$ and $y_i$ respectively, then $y_i$.

![Figure 3. The learned projection matrix and the corresponding groundtruth in the simulation experiments. The white parts are zeros and the black parts are non-zeros.](image-url)
are generated as $y_i = X_i \tilde{w}^T_i + \epsilon_i$ where $\epsilon$ is the noise vector drawn from normal distribution.

We verify the effectiveness by applying our approach on the dataset and comparing the learned projection matrix with the groundtruth. Figure 3 shows one example of the learned projection matrix 3(b) with the comparison of the groundtruth 3(a) where the white part represents zeros and the black part represents non-zeros. The result shows that our approach is able to correctly capture the correct group sparse structures.

### 4.2. Real Data Experiments

We compared our approach with three regularization based feature selection approaches:

**No sharing** Each attribute is treated as a single cluster by extending the lasso regularizer [24] into multi-task learning framework. Specifically, the approach can be formulated as:

$$\text{arg min}_W \mathcal{L}(W|X,Y) + \rho \sum_{i=1}^{m} \|w_i\|_1 \quad (15)$$

**All sharing** All attributes are treated as a whole group by utilizing an $\ell_{2,1}$ norm for feature selection [1].

**Off-line grouped** Attributes are grouped by off-line k-means clustering for feature selection [13]. Since the group information may not be always available, for a fair comparison, we first learn a model $W_0$ independently for each attribute by logistic regression. Then k-means is imposed on $W_0$ to acquire an estimation of the attribute cluster followed by the group feature selection of [13].

For all approaches, the super parameters are selected via cross-validation. We cannot get the number of cluster $k$ without any prior knowledge, thus we also select $k$ by the prediction accuracy on a small subset of datasets.

The experiments are conducted on three benchmark datasets: aYahoo [7], Animals with Attributes (AwA) [18] and SUN attribute [20] datasets. All the datasets are standardized to zero-mean and normalized by the standard deviation. Some attributes are intentionally eliminated if the label is extremely unbalanced, e.g., only less than 0.1% are labeled as 0 or 1. We set 0.5 as a threshold if only continuous attribute labels are given. The statistics of the data are summarized in Table 1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>aPascal/aYahoo</th>
<th>AwA</th>
<th>SUN</th>
</tr>
</thead>
<tbody>
<tr>
<td># of images</td>
<td>15339</td>
<td>30475</td>
<td>14340</td>
</tr>
<tr>
<td># of attributes</td>
<td>64</td>
<td>85</td>
<td>102</td>
</tr>
<tr>
<td># of classes</td>
<td>32</td>
<td>50</td>
<td>611</td>
</tr>
<tr>
<td># of features</td>
<td>2429</td>
<td>1200</td>
<td>1112</td>
</tr>
</tbody>
</table>

Table 1. Statistics of real world dataset.

To obtain a good representation of the high-level attributes, we require that the features can capture both the spatial and context information. Thus, we constructed the features by pooling a variety types of feature histograms including GIST, HoG, SSIM.

#### 4.2.1 Attribute Prediction Accuracy

We first compare our proposed approach with the baselines on the attribute prediction accuracies. For aPascal/aYahoo and AwA datasets we use predefined seen/unseen split published with the datasets. For SUN dataset, 60% of categories are randomly split out as “seen” categories in each round with the rest as “unseen” categories. During training 50% of samples are randomly and carefully drawn from each seen categories to ensure the balance of the positive and negative attribute labels. The rest samples from “seen” classes and all samples from “unseen” classes are used for testing.

**Overall Accuracy** Table 2 shows the average prediction accuracy of each approach over all attributes by running the experiment 10 rounds. The result shows that for both “seen” and “unseen” categories, the “no-sharing” and “all-sharing” approaches generate similar accuracies, which are both less than the “off-line grouped” approach. Our proposed approach further outperforms the “off-line grouped” approach by $2\% \sim 4\%$. The “off-line grouped” approach employs more correlation during learning than “no-sharing”, and decorrelates low-correlated attributes by off-line clustering compared with “all-sharing”, thus achieves better prediction performance. However, the off-line K-means usually can not obtained an optimized clustering structure due to the high-dimensional data, thus usually is able to achieve a sub-optimized result. Our approach iteratively optimizes the clustering structure and the projection model, which achieves the best performance.

Figure 4 illustrates the prediction accuracies when only 20 attributes of AwA datasets are involved for learning. The result shows that considering each attribute, our approach still outperforms the baseline approaches most of the case.

**Human v.s. Machine** We also compare the proposed approach with the human-defined semantic groups. Following [13], the experiments are conducted on aPY-25 with 25 attributes in 3 groups, and AwA with 81 attributes in 9 groups. On aPY-25, the proposed approach achieves accuracies of 64.24\%\%60.03\% on “seen” \“unseen” categories, with the comparison of 64.20\%\%60.07\% achieved by human-defined groups. On AwA, the accuracies achieved by the proposed approach are 62.54\%\%58.37\% compared with 62.50\%\%58.34\% by human-defined groups. The result shows that our proposed approach is able to achieve comparable performance with human-defined approach.

**Learning curves** We then explore the learning curves of the prediction performance with respect to the number of attributes and training samples on three datasets, which are shown in Figure 5. In this experiment we intentionally leave out a small portion of data for training to observe
which decreases the performance gain. Knowledge but also noise during learning for all approaches, be shared. More shared information would bring additional group structure contributes more when less information can be utilized for learning at the first phrase. However, with the increase of the attribute numbers, noises and weak-correlation are also mis-used during learning, which cause the prediction accuracies drop down in the second phrase. Since our proposed approach captures the correlation based on grouped structure, larger performance gains are achieved when the number of attributes increases, e.g., on aYahoo dataset 2% performance gain is achieved with 5 attributes while 4% is achieved with 19 attributes, compared with “all sharing” approach.

The right three plots in Fig. 5 show the learning curves with different number of attributes involved in the learning task. Among all approaches, the prediction accuracies first go up with the increase of attribute numbers, and then drop down. This means that the more number of attributes, the more information can be utilized for learning at the first phrase. However, with the increase of the attribute numbers, noises and weak-correlation are also mis-used during learning, which cause the prediction accuracies drop down in the second phrase. Since our proposed approach captures the correlation based on grouped structure, larger performance gains are achieved when the number of attributes increases, e.g., on aYahoo dataset 2% performance gain is achieved with 5 attributes while 4% is achieved with 19 attributes, compared with “all sharing” approach.

The right three plots in Fig. 5 show the learning curves with different number of training samples involved. The result shows that all approaches achieve higher performance with the increase of training samples, while the performance gain achieved by the proposed approach goes down. That means the “correct” amount of correlation captured by group structure contributes more when less information can be shared. More shared information would bring additional knowledge but also noise during learning for all approaches, which decreases the performance gain.

Case demonstration For an intuitive understanding, we illustrate some success and failure attribute prediction cases of the proposed approach.

Figure 6(a) left three columns illustrate some examples that our approach successfully predicts the attributes while the “no sharing” approach fails, like “Eating”, “Wood” and “Sport”. In such cases, the attributes are either usually not very obvious, e.g., the third image in the first row is easily confused with general “traffic” scene, or merely appear in a small portion of the image, e.g., the “Eating” attribute of the first image in the second row mainly reflected by the pizza in the left corner. Such attributes are not easily detected purely on visual image but can be implied by the correlation from other attributes, e.g., the attribute “wood” usually coexists with some other attributes like “table” in the “dorm” scene. The right two columns demonstrate examples that our approach fails but “no sharing” succeeds. The main reason is that the content of the image is dominated by some other objects, like the buildings in the forth columns, which may be predicted as “no flower”, influenced by some other attributes frequently associated with building images.

Figure 6(b) left three columns give some example that the proposed approach successfully predicts the attributes but failed by the “all sharing” approach. By learning all attributes together, the prediction is easily degraded by instances having weak correlation, e.g., farms should have attribute “grass”. The proposed approach may filter such inappropriate information by partition attributes into groups. The right two columns show some failure examples of our approach but succeed in “all sharing”. Based on our observation, in such cases the objects reflecting the attributes are usually blurred (the fifth image in row one) or in a very small region (the fourth image in row one), which is hard.
Figure 5. Learning curves of prediction accuracy corresponding to the number of attributes and training samples.

4.2.2 Zero-shot Learning

We also experimented on the zero-shot learning problem on all three datasets. Zero-shot learning aims to learn a classifier based on training samples from some seen categories, and classify some new samples to a new unseen category. We adopt the Direct Attribute Prediction (DAP) framework proposed in [18] with attribute prediction probability from each approaches as input. Since only continuous image level attribute labels are provided on the SUN dataset, we construct the class level attribute labels by thresholding the average attribute label values of all samples from the class. Same “Seen” “Unseen” categories splits are adopted as previous experiments. Average classification accuracies of 10 rounds experiment are reported in Table 3. The result shows that on aYahoo and AwA, our approach achieves significant performance gains than the baseline approaches.

![Image illustration of prediction results.](image-url)

(a) Some successful (left three columns) and failed (right two columns) of the proposed approach compared with no sharing.

(b) Some successful (left three columns) and failed (right two columns) of the proposed approach compared with all sharing.

Table 3. Zero-shot learning accuracy on both real dataset.

<table>
<thead>
<tr>
<th></th>
<th>aYahoo</th>
<th>AwA</th>
<th>SUN</th>
</tr>
</thead>
<tbody>
<tr>
<td>No sharing</td>
<td>0.1822</td>
<td>0.2945</td>
<td>0.1866</td>
</tr>
<tr>
<td>All sharing</td>
<td>0.1834</td>
<td>0.2953</td>
<td>0.1842</td>
</tr>
<tr>
<td>Off-line Grouped</td>
<td>0.2052</td>
<td>0.3085</td>
<td>0.2010</td>
</tr>
<tr>
<td>Proposed</td>
<td><strong>0.2262</strong></td>
<td><strong>0.3258</strong></td>
<td><strong>0.2133</strong></td>
</tr>
</tbody>
</table>

The large number of categories in SUN dataset make the classification problem very hard which leads to all low performance of all approaches. Our approach still works better than the baseline approaches.

5. Conclusions

In this paper, we proposed a clustered multi-task feature learning framework for semantic attribute prediction. Our approach employs both K-means and group-sparsity regularizers for feature selection. The K-means regularizer partitions the attributes into different groups where strong correlation lies among attributes in the same group while weak correlation exists between groups. The group-sparsity regularizer encourages intra-group feature-sharing and inter-group feature competition. With an efficient alternating optimization algorithm, the proposed approach is able to obtain a good group structure and select appropriate features to represent semantic attributes. The proposed approach was verified on both synthetic and real image datasets with comparison with state-of-the-art approaches. The result shows effective group structure identification capability of our method, as well as its significant performance gains on both attribute prediction accuracy and zero-shot learning classification accuracies.
References


