Augmenting the semantic attribute representation with the discriminative features has been proved to be an effective method for improving the performance of object classification. However, how to make the expanded features more effective and discriminative is still an open problem. In this paper, we propose a Sequential Augmented Features Learning method (SAL) to implement semantic attribute augmentation. In our SAL method, the augmented non-semantic features are learned one by one under a sequential error-correcting scheme so that we can obtain more discriminating power with very compact expanded features. Extensive experiments are conducted on a public dataset and the results show that our approach achieves encouraging performance.

Index Terms — augmented feature, sequential feature learning

1. INTRODUCTION

In recent years, learning representations in terms of semantic attributes [1] offers more detailed mid-level representation that can be both detected by machines and nameable by humans. Because attributes information can be shared across the categories, they play an important role in zero-shot learning of a new category which only has a list of attributes description [2].

However, in many previous works, the semantic attributes are usually pre-defined and quantitatively limited, which may not be always sufficient enough to fulfill classification for all classes. Taking Fig. 1 for example, it is difficult to distinguish the differences between cats and dogs purely with some pre-defined semantic attributes (e.g., “furry”, “tail”, and “claws”), even though they are from different categories and show many obvious dissimilarities. Therefore, the work [3] extends the semantic attribute representation with additional mid-level features, which are inferred by an autoencoder in combination with the folk wisdom criterion (e.g., “stay close to your friends and run away from your enemies”). However, they can not learn the new mid-level features based on the already learned ones. If the number of the additional mid-level features is changed, they have to learn the encode function again which is high time complexity. In [4] some special designed attributes, for example in categorization case discriminative attributes (e.g. “cats and dogs have it but sheep and horse do not have”), have been introduced and added to the traditional attributes list. However, the work wants to keep newly learned attributes possessing both semantic and discriminative capabilities. Unfortunately, the attempt is often achieved either with high time complexity or in sacrifice of the performance.

In this paper, we propose a novel Sequential Augmented Feature Learning (SAL) method to improve the attribute-based classification performance. As demonstrated in Fig. 1, our new image representation consists of two parts: pre-defined semantic attribute part and augmented non-semantic feature part. The former part still keeps the traditional semantic properties of attributes. The augmented part whose name is not known enhances the discriminative ability of the integrated attribute representation. Each augmented feature is built upon the low-level features (e.g. color and texture features) with a learner which is computed based on the observation that two instances should probably share the same...
(different) augmented feature if they are from the same (different) class. Moreover, in our proposed method, each augmented feature learner is obtained in a sequential manner such that the errors made by the previous learner can be corrected by the following one. Therefore, the attributes expanded by our approach have more powerful discriminating ability. Extensive experiments verify the promising performance of the proposed method.

2. ALGORITHM

Assume $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{d \times N}$ to denote $N$ images with $d$-dimensional features, and the semantic attributes $a_i = W_A^T x_i \in \{-1, 1\}^n$ can be obviously inferred with the already learned semantic attribute predictors $W_A \in \mathbb{R}^{d \times n}$ ($n$ is the number of semantic attributes). We assume the labeled images denoted by $X_l \subset X$, $X_l \in \mathbb{R}^{d \times l}, l \ll N$.

We sequentially parameterize the augmented feature learners $W_B \in \mathbb{R}^{d \times m}$ ($m$ is the number of augmented features) and each column of matrix $W_B^T, k \in \{1, \ldots, m\}$ is a learner for studying. With hybrid representation $[a_i^T b_i^T]^T$, where $b_i = W_B^T x_i$, we hope that our proposed method can obtain better classification performance than the methods with the semantic attributes representation alone. In our work, we simply consider binary representations for the semantic and non-semantic representation space (e.g., $b_i \in \{-1, 1\}^m$). Nearest-neighbor based object categorization method is used after the hybrid representation is inferred.

2.1. The Problem Formulation

Our method aims to parameterize feature learners $W_B$ that minimize the error on the labeled training data $X_l$, at the same time maximize information from each dimension of the augmented feature like in the work [5]. To choose a good $W_B$ an objective function measuring the empirical accuracy on the labeled data can be defined as:

$$
\mathcal{L}(W_B) = \sum_{ij} S_{ij} \{\text{sgn}([W_A W_B]^T x_i)]^T \{\text{sgn}([W_A W_B]^T x_j)]\},
$$

where $S_{ij}$ is the number of semantic attributes. We assume the labeled images denoted by $X_l \subset X$, $X_l \in \mathbb{R}^{d \times l}, l \ll N$.

Here we use the category information of the labeled samples to initialize the original matrix $S \in \mathbb{R}^{l \times l}$; given a labeled image set $X_l$ if image $x_i$ and $x_j$ share the common label then we set $S_{ij} = 1$ and we want to learn $W_B$ that gives the same features when we maximize the objective function. While image $x_i$ has a different label from $x_j$ then we set $S_{ij} = -1$ and if $i = j$ then $S_{ij} = 0$. Since $W_A$ is already learned and known from a separate method for attribute-prediction, the objective function in Eq. (1) can be further represented as,

$$
\mathcal{L}(W_B) = \sum_{ij} S_{ij} |\text{sgn}(W_B^T x_i)^T |\text{sgn}(W_B^T x_j)|
$$

Because the object function is non-differentiable, we replace the sign of feature values with their signed magnitude as follows:

$$
\mathcal{L}(W_B) = \sum_{ij} S_{ij} W_B^T x_i^T W_B^T x_j
$$

and the above function can be written in a matrix form as

$$
\mathcal{L}(W_B) = \frac{1}{2} \text{tr}\{W_B^T X S X^T W_B\}
$$

However, in a small shot learning scenario maximizing the objective function may lead to severe overfitting. We propose to maximize the variance of the augmented feature representations of the whole dataset $X$ for the sake of maximizing information provided by each dimension of the augmented feature. The regularization part can be formulated as,

$$
\mathcal{J}(W_B) = \text{Var}(W_B^T X) = \sum_{k} E[\|W_B^T X - E(W_B^T X)\|^2]
$$

$$
= \frac{1}{N} \text{tr}\{W_B^T X X^T W_B\}
$$

here we have used the properties that the data is zero-centered, i.e., $E(W_B^T X) = 0$.

Combining Eq. (4) and Eq. (5), we maximize the following regularized objective function for learning the augmented feature representations

$$
\mathcal{L}(W_B) = \mathcal{J}(W_B) + \lambda \mathcal{L}(W_B)
$$

$$
\propto \text{tr}\{W_B^T X X^T W_B\} + \lambda \text{tr}\{W_B^T W_B\}
$$

and

$$
L = XX^T + \lambda XX^T + XX^T
$$

The learning of the optimal $W_B$ becomes a typical eigenproblem. We learn the predictors $W_B^k, k = 2, \ldots, m$ iteratively: first, the matrix $S$ is updated by imposing higher weights on the image pairs on which augmented feature wrongly assigned by the previous learned non-semantic feature learner; then, we learn the current learner that corrects the most errors made by the previous one from the top $p$ eigenvectors of the current matrix $L$ [6, 7].

We define

$$
Q^k = X_l^T W_B^k (W_B^k)^T X_l,
$$

which simply measures the sign magnitude of pairwise relationships of the k-th augmented feature value on $X_l$. If $Q^k_{ij} > 0$, it means that both image $x_i$ and $x_j$ share the common response sign of k-th augmented feature. If $Q^k_{ij} < 0$, it means that only one image have this feature. The magnitude of $Q^k_{ij}$ reflects the degree of its confidence. We can obtain the updating rules for $S$ as follows,

$$
S_{ij}^{k+1} = \{\begin{array}{ll}
S_{ij} - Q^k_{ij}, & \text{sgn}\{S_{ij} \cdot Q^k_{ij}\} < 0 \\
0, & \text{sgn}\{S_{ij} \cdot Q^k_{ij}\} \geq 0
\end{array}
$$

where $\text{sgn}\{S_{ij} \cdot Q^k_{ij}\} < 0$ means the images from the same class ($S_{ij} > 0$) have different responses by the k-th feature.
Algorithm 1 The Proposed SAL Algorithm

Input:
Image set $X \in \mathbb{R}^{d \times N}$ with attribute predictors $[W^1_A, \ldots, W^m_A]$, labeled image set $X^l_i \in \mathbb{R}^{d \times t}$ and length of augmented feature dimension $m$; Parameters constant $\lambda$.

Output:
Augmented feature learners $[W^1_B, \ldots, W^m_B]$.

1: Initial similarity matrix $S^1$;
2: for $k = 1$ to $m$ do
3: Compute $L^k$ as follows,
$$L^k = X X^T + \lambda X_i S^k X_i^T$$
4: Extract the top $p$ eigenvectors of $L^k$
5: Select the eigenvector with the highest score as $W^k_B$
6: Update the similarity matrix $S^k$ by Eq. (9)
7: Compute the residual:
$$X = X - W^k_B (W^k_B)^T X$$
8: end for

learner (one of them has the positive response and the other not), or the images from the different classes ($S_{ij} < 0$) have the same response. For this case, we increase the penalize of the relevant image pairs. From Eq. (9) we can see that we only change the magnitude of $W_{ij}$. If $S_{ij} < 0$, we update the similarity matrix $S^k$ by Eq. (9) as follows:

$$S^k_{ij} = S^k_{ij} (S_{ij} \cdot Q^k_{ij} < 0) \odot (S_{ij} \cdot Q^k_{ij} > 0)$$

where $\odot$ denotes the element-wise multiplication and $Q^k_{ij} = X_i^T W^k_B (W^k_B)^T X_i$. The score reflects how many wrong feature assigned pairs by the $k$-th predictor are corrected by the learner $W^k_B$. We select the eigenvector with the highest score as $W^k_B$. In our experiment $p$ is set to 5.

Note before we learn $W^k_B$ the matrix $S$ is updated for the first time according to the neighborhood relationship with images’ semantic attribute representation $a_i, i = 1, \ldots, l$. If any relationship is wrong inferred compared with the ground truth, corresponding $S_{ij}$ will be adjusted like the scheme described above.

The detail of our proposed algorithm is shown in Algorithm 1.

3. EXPERIMENTS

In this section, in order to validate the classification effectiveness of the our proposed SAL algorithm for object categorization, extensive experiments are conducted on a publicly available dataset by comparing with one baseline and the state-of-the-art algorithms. We use 3-nearest neighbors rule for pattern classification. The parameter $\lambda$ is selected by a cross validation model. The compared scheme is listed as follows.

- original: The baseline method using the original low-level features as image representation.
- attribute: The approach proposed in [2] which only uses pure array of semantic attributes as image representation.
- SALnaive: The batch mode solution for our proposed method which is the naïve SAL without sequential selecting the augmented feature learner.
- SAL: Our proposed method.

3.1. Dataset

In this section, we conduct experiments on the Animals with Attributes (AwA) dataset [2] to evaluate the performance of the proposed approach.

AwA dataset: It consists of 50 object categories and 30,475 images in total. The original low-level features we used here are SURF descriptions provided by the dataset. The dataset also provides an 85-dimensional semantic attribute vector. We follow the works [2] and [3] to design our experimental protocol. We use a $\ell_2$-regularized $\ell_2$-loss SVM model with 40 classes images to learn the semantic attribute predictors matrix $W_A$. 1,500 images, i.e. 150 per class from the rest 10 classes are used as the test set. The other images in this 10 classes are used to learn the augmented feature learners $W_B^k$.

We repeat the entire training and test process 5 times to get better statistics of the performance.

3.2. Results on Animals with Attributes dataset

Note that in the Animals with Attributes dataset the instances in the same class have the same attributes. We randomly label 2,4,6,8,10 samples per class for training the augmented feature learners. Because in this dataset the SURF feature is extracted over the whole image the background probably affects the attributes prediction of the foreground. We wish to see that after we add the augmented feature the discriminative performance of hybrid representation can be improved. So we set the number of dimension $m = 10, 20, 50, 85$ respectively. The results are illustrated in Fig. 2.

From the results we can observe that the semantic attribute representation have better performance than the original feature. The performance of the method SALnaive is compa-
Fig. 2: Augmented feature representations using proposed sequential approach. Performance of the categorization across the different dimensions ($m = 10, 20, 50, 85$).

Fig. 3: The accuracy of sequential approach training with different number (2, 4, 6, 8, 10 respectively) of labeled samples. Mean accuracy over 5 runs.

In this paper we propose a sequential augmented feature learning method to improve the classification performance of the semantic attributes representation. For each bit augmented feature, its learner is computed to reflect its discrimination and maximizes the information. In our proposed approach each augmented feature learner is learned to correct the mistakes which are made by the previous learner as far as possible. Extensive experiments show that after we add the augmented feature the discriminative performance of hybrid representation has been improved. The proposed method achieves encouraging results.

Acknowledgements. This work was supported in part by the 973 Program under Project 2010CB327905, by the National Natural Science Foundation of China under Grant No. 61170127, 61070104 and Key Project of Chinese Academy of Sciences under Grant No. KGZD-EW-103-5.

5. REFERENCES