SHARING MODEL WITH MULTI-LEVEL FEATURE REPRESENTATIONS

Li Shen[1], *Gang Sun*[1,2], *Shuhui Wang*[3,*], *Enhua Wu*[2,4], *Qingming Huang*[1,3]

[1] University of Chinese Academy of Sciences, Beijing, China
[2] State Key Lab. of Computer Science, Inst. of Software, CAS, Beijing, China
[3] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[4] University of Macau, Macao, China

ABSTRACT

Hierarchical classification models have been proposed to achieve high accuracy by transferring effective information across the categories. One important challenge for this paradigm is to design what can be transferred across the categories. In this paper, we propose a novel method to learn a sharing model by taking advantage of multi-level feature representations. Unlike many of the existing methods which learn the sharing model based on identical feature space, multi-level feature detectors enable our model to capture rich visual information in hierarchical category structure. Moreover, hierarchical classifier parameters associated with multi-level feature representations are learned to model the visual correlation in the hierarchy. The experimental results on Caltech-256 dataset and ImageNet subset demonstrate that our method achieves excellent performance compared with some state-of-the-art methods, and shows the advantage of multi-level information transfer.

Index Terms— Sharing model, multi-level feature representations, object categorization

1. INTRODUCTION

Visual classification with many classes is one of the core problems in computer vision, and poses significant challenges. Due to the imbalance of class distribution, the performance of classification model on rare categories may be limited by insufficient training data. It is nature to cluster similar categories into groups, and generate a tree structure [1, 2]. Based on the tree-shaped hierarchy, the classification models for multiple classes can be trained jointly which enables rare category to benefit from other related categories. Many methods have been developed to share various information (e.g., a global prior [3, 4], statistical parameters [5, 6, 7, 8]) across multiple classes, and most of them train the hierarchical models on identical feature space. However, designing what can be transferred across categories is a critical issue, which is worth exploring for sharing model learning.

In this paper, we propose a novel approach which takes advantage of multi-level feature representations to learn a sharing model. As shown in Fig. 1(b), multi-level feature detectors (filters) can capture various patterns from image data, from simple oriented edges, to mid-level patterns indicative of object parts, and high-level ones indicative of objects. Our main contribution is to exploit these features to learn a hierarchical sharing model which can transfer multi-level information among the classes. As shown in Fig. 1(a), the model parameters learned in the hierarchy are associated with the feature detectors of different levels. For the root (global) node, global parameters (w^0) are associated with the representations generated from global feature detectors. The group parameters (w_1^1, w_2^1) are learned to select the features which are useful to describe the common information in the group, and the parameters can be shared by child classes. For example, *leg* is usually captured in {*horse, dog*}, and *wheel* tends to appear in {*car, motorbike, tricycle*}. The two features are corresponding to large weight in group *animal* and *vehicle* respectively. Specific parameters ($w_1^2, w_2^2, ...$) are associated with high-level feature detectors to discover more class-specific information towards classes. The experimental results demonstrate that our method benefits from multi-level feature representations and achieves excellent performance.

2. MULTI-LEVEL FEATURE LEARNING

Feature learning aims to extract and organize discriminative information from data. Multi-level feature detectors learned with deep architecture [9, 10, 11, 12, 13] have great advantages over the ones extracted by feature engineering with shallow structure [14, 15, 16, 17, 18, 19]. In deep architecture, simple features are learned first, and then more complex features are built up by composing the simpler ones together. To deal with realistic-sized (e.g., 200×200 pixel) images, we leverage convolutional network architecture [9, 11] to generate representations. The architecture is composed of multiple

*Corresponding author.

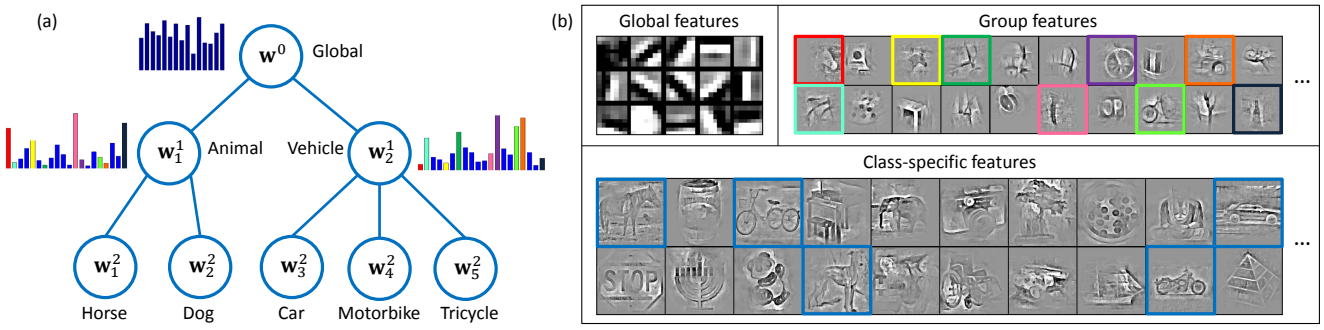

Fig. 1. Overview of our approach. (a) Sharing model with multi-level feature representations. Global parameters (w^0), group parameters (w_1^1,w_2^1), and specific parameters ($w_1^2, w_2^2, ...$) are learned in the hierarchy to capture information from multi-level features. (b) Visualization of the feature detectors in each layer. The colorful histograms in (a) illustrate the group parameters, and the colors are corresponding to the box colors on group features in (b). For internode *Animal*, the features indicative of animal parts are associated with large weight. For *Vehicle*, the features captured *wheel* are associated with large weight. The specific parameters on the five leaf nodes have maximum value on the class-specific features with blue boxes in (b) respectively.

alternating layers of *convolution* and *pooling* operator.

Convolution: In convolutional layer, we apply convolutional sparse coding (deconvolution) [9] which attempts to minimize the reconstruction error of the input image on an over-complete set of feature maps. Consider the first layer, for input image x, the reconstruction can be formulated as:

$$L(x, v, f) = \frac{1}{2}\|x - \sum_{k=1}^{K_0} f_{k,0} * v_{k,0}\|^2 + \sum_{k=1}^{K_0} |v_{k,0}|, \quad (1)$$

where $*$ is the convolution operator. $f_{k,0}$ denotes the $m \times m$ feature detector (filter) common to all the images. $v_{k,0}$ denotes the feature map specific to each image. K_0 is the number of filters. $\hat{v} = \arg\min_v L(x, v, f)$ is a unique solution, which can be gained based on the sparsity constraint on v.

Pooling: In pooling layer, max-pooling operates in local neighborhood to shrink the outputs of the convolutional layer. Specifically, each unit in pooling layer computes the maximum (absolute) value in a small region of feature maps, and the locations. The outputs (pooled maps p and locations s) are used in the next convolutional layer. This operator enables the representations of subsequent layers to be invariant to small translation, and to capture the patterns at a larger scale.

Based on the hierarchical feature learning, multi-level visual patterns can be captured by the global features, group features as well as specific features. As shown in Fig.1(b) (visualization according to the strategy in [9]), simple patterns can be firstly captured, such as orientated edges, which can be detected in all the categories. Thus, they can be regarded as global features shared by all the classes. Mid-level patterns indicative of object parts (e.g., *wheel* and *animal head*) can be discovered, which have significant statistical distributions, such as *wheel* can be frequently found in *vehicle* and *animal head* tends to appear in *animal*. These features can be shared by a group of related classes as group-based properties. Moreover, high-level patterns describe specific object classes which can give the object an overall description. We employ another spatial pyramid pooling [16] on feature maps to generate feature vectors, which represent the statistical distribution of the features in an image. The generated feature vectors are denoted with z^0, z^1 and z^2.

3. LEARNING SHARING MODEL WITH MULTI-LEVEL REPRESENTATIONS

3.1. Traditional Separate Classification Model

Consider a classification problem with $D = \{(x_i, y_i)\}_{i=1}^N$ of N labeled training images. Each example belongs to one of T classes, $y_i \in \{1, 2, ..., T\}$. In the standard classification model learning, the problem can decompose to T subproblems, which amounts to T separate binary (one-versus-all) classification models without sharing any information between them. Assuming that these models are trained based on the feature set $Z \subset \mathbb{R}^m$:

$$\min_W \sum_{t=1}^{T} \left(\ell(D, Z, w_t, \mathcal{F}_t) + \frac{\lambda}{2}\|w_t\|^2\right), \quad (2)$$

where w_t denotes parameter vector corresponding to the t-th classification model, and W denotes the parameter matrix composed of w_t as columns. $\|w_t\|^2$ is the regularization term, which can be regarded as Gaussian prior over model parameters. $\ell(D, Z, w_t, \mathcal{F}_t)$ denotes the loss function on the t-th class with regard to discriminative function $\mathcal{F}_t(w_t, z_i) = w_t^T z_i$, where z_i denotes the feature representation of sample x_i. The function can be defined with hinge loss:

$$\ell_{hinge} = \sum_{i=1}^{N} \max\left(0, 1 - \mathcal{C}_t(y_i)\mathcal{F}_t(w_t, z_i)\right), \quad (3)$$

or probabilistic log-loss:

$$\ell_{prob} = \sum_{i=1}^{N} \log\left(1 + e^{-\mathcal{C}_t(y_i)\mathcal{F}_t(w_t, z_i)}\right), \tag{4}$$

where $\mathcal{C}_t(y_i) = 1$ if $y_i = t$ and -1 otherwise.

3.2. Hierarchical Model with Multi-level Representations

Given the hierarchical structure, the categories are grouped to G hyper-classes. For example, *horse* and *dog* belong to the group which represents *animal*, whereas *car*, *motorbike* and *tricycle* belong to the hyper-class of *vehicle*. In our sharing model, each node in the hierarchy is associated with a separate parameter vector. The parameter vectors in different layers are associated with the representations of different levels. For example, given the parameters of class *horse*, $w_1^+ = \{w^0, w_1^1, w_1^2\}$, for each example x_i, the discriminative function in Eq. 2 can be rewritten by using multi-level representation $z_i^+ = \{z_i^0, z_i^1, z_i^2\}$:

$$\mathcal{F}_1(w_1^+, z_i^+) = {w^0}^T z_i^0 + {w_1^1}^T z_i^1 + {w_1^2}^T z_i^2, \tag{5}$$

where w_0, w_1^1, w_1^2 represent the global, group, and specific parameter vectors for the class respectively. When the multi-level representation z_i^+ is replaced by identical feature z_i, the Eq. 5 reduces to $\mathcal{F}_1(w_1^+, z_i) = (w^0 + w_1^1 + w_1^2)^T z_i$, which is similar to the formulation in [7]. Compared with the sharing model in [7], our method combines more statistical information with the aid of multi-level representations.

Drawing on connection to the standard classification model, the learning problem can be formulated as following:

$$\begin{aligned} \min_{W^+} \sum_{t=1}^{T} \ell(D, Z^+, w_t^+, \mathcal{F}_t) + \frac{\lambda_0}{2}\|w^0\|^2 \\ + \frac{\lambda_1}{2}\sum_{g=1}^{G}\|w_g^1\|^2 + \frac{\lambda_2}{2}\sum_{t=1}^{T}\|w_t^2\|^2, \end{aligned} \tag{6}$$

where $W^+ = \{W^0, W^1, W^2\}$, W^0 only contains w^0 as one column. W^1 is comprised of w_g^1 as column vectors, and W^2 denotes the matrix composed of w_t^2. The loss function $\ell(*)$ can be defined by Eq. 3 or Eq. 4.

Given the tree structure, the model parameters can be optimized efficiently using an iterative procedure [7], as shown in Algorithm 1. The target function in Eq. 6 can be decomposed into several sub-problems. For example, when W^0 and W^2 are given, W^1 can be optimized efficiently as a traditional classification model.

Algorithm 1: Model Parameter Optimization

Input : Tr: the tree structure, G: the number of groups, T: the number of categories
Output: W^0, W^1, W^2

```
Initialize: W^0 = O, W^1 = O, W^2 = O
while not converged do
    Given W^1 and W^2, optimize global parameters W^0 using Eq. 6.
    for g = 1 to G do
        Given W^0 and W^2, optimize group parameters w_g^1 using Eq. 6.
    end
    for t = 1 to T do
        Given W^0 and W^1, optimize specific parameters w_t^2 using Eq. 6.
    end
end
```

4. EXPERIMENTAL RESULTS

In this section, we evaluate our method on two datasets: Caltech-256 [20] and ImageNet subset [21]. Our goal is to 1) verify that our method is more accurate than the separate classification models; 2) show the advantage of learning the sharing model with multi-level feature representations.

Our method is mainly comprised of two aspects : hierarchical sharing model and multi-level feature learning. In order to evaluate the performance of our method, we compare it with the following methods:

1. Flat Model + ScSPM [16] (F-ScSPM). Ignore hierarchical structure and the model learning is based on flat structure. The classes are trained separately with one-versus-all strategy, as described in Section 3.1. The feature is computed by sparse coding.

2. Sharing Model + ScSPM (S-ScSPM). The model is learned based on hierarchical structure [7] with ScSPM.

3. Flat Model + Multi-Level Feature Representations (F-MLF). The model is based on flat structure. The multi-level feature representations are learned as described in Section 2.

Each image has be converted to gray-scale and resized to 150×150, using zero padding to preserve the aspect ratio. For parameter details, dense SIFT [22] is used for coding in ScSPM, and dictionary size is 1024. For model learning, hinge loss is applied, as shown in Eq. 3. The regularization parameters λ in Eq. 6 are set to one. Multi-level feature learning is followed the configuration in [9]. The hierarchical models are evaluated on a pre-computed tree structure of 3 layers. The structure is constructed followed the strategy in [23].

4.1. Caltech-256

Caltech-256 [20] is a standard multi-class object recognition dataset, which is comprised of 256 categories. For each class, we randomly sampled 60 images, and split them into one half (30) as training data and the other half (30) for testing. Table 1 shows the accuracy of different methods in Caltech-

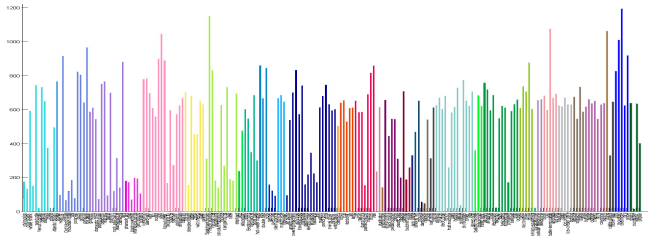

Fig. 2. The distribution of the samples for 200 concepts selected from ImageNet.

Table 1. Accuracy of four methods in Caltech-256 and ImageNet subset.

Algorithm	F-ScSPM	S-ScSPM	F-MLF	Ours
Caltech-256	29.7%	32.4%	32.8%	**34.6%**
ImageNet	18.6%	21.4%	20.7%	**23.9%**

256. Our method achieves better results compared with the baseline methods. The result shows the advantage of incorporating category hierarchy with multi-level feature representations. Compared with the methods without multi-level feature learning (F-ScSPM, S-ScSPM), multi-level feature learning can capture rich visual information with multiple scales, which is helpful to improve performance. On the other hand, the performance difference between F-MLF and our method shows the sharing model learning can effectively exploit and transfer information with the aid of multi-level visual feature representations. This can be shown in Fig 1(b). Mid-level features capture the patterns indicative of object parts, the difference between the value of group parameters shows the bias towards these features, that is to say, different visual information can be shared between related classes. The features with large weight reveal the common information in the group, such as the pattern like *wheel* towards the group which contains *car*, *motorbike* and *tricycle*.

4.2. ImageNet dataset

ImageNet [21] is a large scale dataset where the class concepts are organized based on WordNet [24] structure. We randomly select 200 categories which covers wide domains of semantics, such as animal, plant, container, sport. The number of class samples is quite different, from several to thousands, as shown in Fig. 2. We split the samples of each class into two equal sets: one is for training and the other is for testing. Table. 1 shows the results achieved by all the methods on these categories. We can observe that the sharing models (S-ScSPM, our method) show better performance compared with the models based on flat structure. Moreover, the increment of accuracy over F-ScSPM (shown in Fig. 3) demonstrates that the model based on sharing paradigm has strong ability when tackling the classes with imbalance distribution of samples. Information sharing in the hierarchy can allow the categories

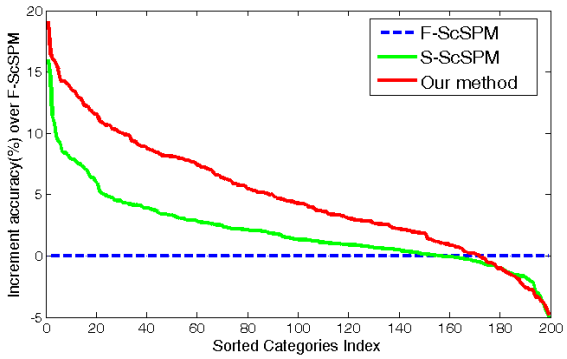

Fig. 3. Improvement of sharing models over F-ScSPM. Categories are sorted by the improvement on accuracy.

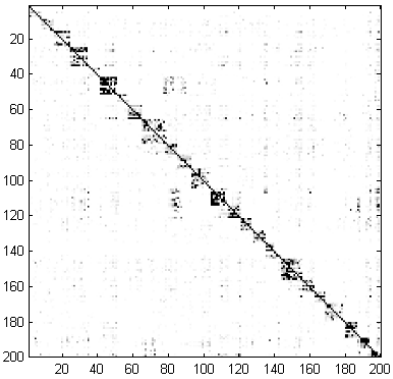

Fig. 4. Confusion matrix of our model on ImageNet subset.

with few samples to borrow statistical strength from related categories. Moreover, as our method exploits rich visual information from multiple levels, it achieves better results compared with S-ScSPM. Fig. 4 displays the confusion matrix of our model on 200 classes. It displays block-structured, indicating that the errors of our model mostly occur on the related classes in a group rather than the arbitrary ones.

5. CONCLUSION

We have proposed to learn a hierarchical sharing model by incorporating with multi-level feature representations. Multi-level visual information is transferred across the related categories in the hierarchy. The experimental results show the effectiveness of our approach. As the hierarchy in our model is pre-computed by some other methods, it may be not optimal for effective information transfer. One future direction is to learn the hierarchy and the model simultaneously.

Acknowledgments This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61025011, 61332016, 61272326, 61303160 and 61390511, and in part by 863 program of China: 2014AA015202, and in part by Research Grant of University of Macau. Li Shen and Gang Sun contributed equally to this work.

6. REFERENCES

[1] A. Zweig and D. Weinshall, "Exploiting object hierarchy: Combining models from different category levels," in *ICCV*, 2007.

[2] T. Gao and D. Koller, "Discriminative learning of relaxed hierarchy for large-scale visual recognition," in *ICCV*, 2011.

[3] T. Evegniou and M. Pontil, "Regularized multictask learning," in *KDD*, 2004.

[4] L. Fei-Fei, R. Fergus, and P. Perona, "A bayesian approach to unsupervised one-shot learning of object categories," in *ICCV*, 2003.

[5] D. Zhou, L. Xiao, and M. Wu, "Hierarchical classification via orthogonal transfer," in *ICML*, 2011.

[6] Z. Kang, K. Grauman, and F. Sha, "Learning with whom to share in multi-task feature learning," in *ICML*, 2011.

[7] R. Salakhutdinov, A. Torralba, and J. Tenenbaum, "Learning to share visual appearance for multiclass object detection," in *CVPR*, 2011.

[8] A. Torralba, K.P. Murphy, and W.T. Freeman, "Sharing visual features for multiclass and multiview object detection," *PAMI*, vol. 29, no. 5, pp. 854–869, 2007.

[9] M. Zeiler, G. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *ICCV*, 2011.

[10] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet classification with deep convolutional neural networks," in *NIPS*, 2012.

[11] H. Lee, R. Grosse, R. Ranganath, and A.Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *ICML*, 2009.

[12] G.E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[13] L. Shen, S. Wang, G. Sun, S. Jiang, and Q. Huang, "Multi-level discriminative dictionary learning towards hierarchical visual categorization," in *CVPR*, 2013.

[14] L. Fei-Fei and P. Perona, "A bayesian heirarchical model for learning natural scene categories," in *CVPR*, 2005.

[15] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering object categories in image collections," in *ICCV*, 2005.

[16] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR*, 2009.

[17] J. Wang, J. Yang, K. Yu, and F. Lv, "Locality-constrained linear coding for image classification," in *CVPR*, 2010.

[18] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *NIPS*, 2009.

[19] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bag of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.

[20] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," Tech. Rep. 7694, California Institute of Technology, 2007.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009.

[22] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," `http://www.vlfeat.org/`, 2008.

[23] N. Zhou and J. Fan, "Jointly learning visually correlated dictionaries for large-scale visual recognition applications," *PAMI*, vol. 99, 2013.

[24] C. Fellbaum, "Wordnet: An electronic lexical database," 1998.