AN ONTOLOGICAL BAGGING APPROACH FOR IMAGE CLASSIFICATION OF CROWDSOURCED DATA

Ning Xu, Jiangping Wang, Zhaowen Wang, Thomas Huang

Beckman Institute, University of Illinois at Urbana-Champaing, USA
{ningxu2,jwang63,wang308}@uiuc.edu, huang@ifp.uiuc.edu

ABSTRACT

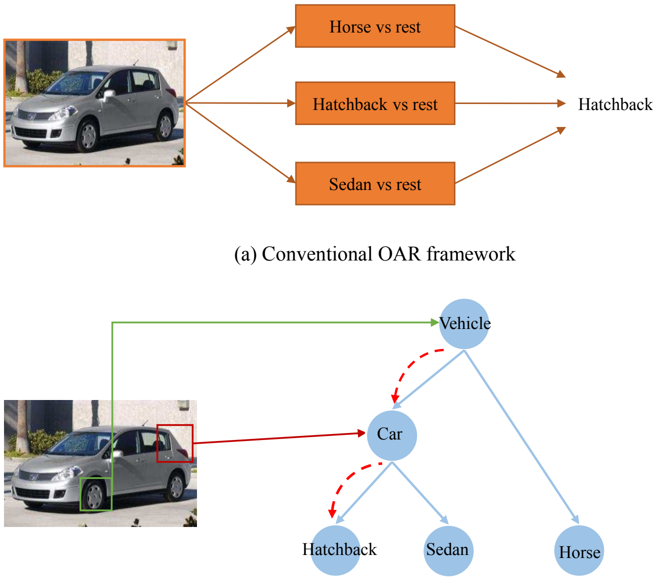
In this paper, we study how to use semantic relationships for image classification in order to improve the classification accuracy. We achieve the goal by imitating the human visual system which classifies categories from coarse to fine grains based on different visual features. We propose an ontological bagging algorithm where most discriminative weak attributes are automatically learned for different semantic levels by multiple instance learning and the bagging idea is applied to reduce the error propagations of hierarchical classifiers. We also leverage ontological knowledge to augment crowdsourcing annotations (e.g., a hatchback is also a vehicle) in order to train hierarchical classifiers. Our method is tested on a vehicle dataset from the popular crowdsourcing dataset ImageNet. Experimental results show that our method not only achieves state-of-the-art results but also identifies semantically meaningful visual features.

Index Terms— Ontology, image classification, hierarchical weak attributes, crowdsourcing

1. INTRODUCTION

Image classification is one of the most important problems in computer vision and pattern recognition. Most existing image classification algorithms [1, 2, 3, 4, 5, 6, 7] treat classes as completely independent both visually and semantically. Features describing the entire images are used to train an one-against-rest multi-class classifier, which can be then applied to test new images against every category exhaustively. However, the omission of inter-class relationships causes the restrictive classification performance especially for categories with fine-grained distinction as well as the high testing complexity when the number of classes grows.

Humans can easily recognize hundreds of thousands of classes since humans use semantic relationships when learning the visual appearance of categories. For example, it's not sensible to remember all the details of "hatchback" when differentiating it from "horse", but more natural to find features corresponding to "car" like "wheel" or "headlight". An ontology is a hierarchical structure consisting of categories and high-level relationships. It provides a potential way to incorporate semantics into the visual recognition system. A simple ontology is shown in Figure 1b.



(a) Conventional OAR framework

(b) The hierarchical framework of our method

Fig. 1: Comparison of conventional image classification algorithms with our ontology-based algorithm. In (a), Conventional methods use the one-against-rest framework and use features about the entire image. In (b), Dotted red lines show that our method classifies categories from coarse to fine grains. Bounding boxes with different colors indicate the learned weak attributes for different semantic levels.

Nowadays the prosperity of World Wide Web makes many large-scale multimedia systems practical with integrated wisdom of the crowds, e.g. the human-annotated large-scale image dataset ImageNet [8]. Often a small number of annotations are labeled for an image. For example, a "hatchback" image is only annotated as "hatchback". However, a "hatchback" can also be labeled as a "vehicle", a "transportation means" or even an "object". Therefore, the crowdsourcing annotations cannot convey the complete semantics. In fact, there is no need to require all possible annotations from crowdsourced data due to the high inefficiency. We can obtain the extra information inexpensively by incorporating ontological prior knowledge. In Figure 1b, we can automatically augment an additional annotation “car” for the hatchback image, while conventional non-ontology-based algorithms have no such advantages.

In this paper, we propose an ontological bagging algorithm to leverage semantic relationships for image classification. In our method, multiple instance learning [5] is used to automatically learn weak attributes for different semantic levels of an ontology. A hierarchical classifier using the learned weak attributes is constructed to classify categories from coarse to fine grains. The bagging approach is applied to further improve the classification performance. Figure 1b illustrate the framework of our method. We evaluate our algorithm on a vehicle recognition dataset from ImageNet [8]. Experimental results show that our method not only achieves the state-of-the-art results but also parallels the human visual system which can locate semantically meaningful image windows for different semantic levels.

The rest of the paper is organized as follows. In Section 2 we will talk about some related work. Section 3 describes our proposed algorithm and experimental results are shown in Section 4. Finally, we conclude the paper in Section 5.

2. RELATED WORK

Image classification has been studied for many years. Most existing algorithms [1, 2, 3] don’t consider inter-class relationships. While sufficient for categories starkly different in visual appearance, they are likely to perform poorly on categories with subtle differences. In this paper, our method leverages ontology to improve image classification accuracy.

Some previous studies [9, 10] have used ontologies for image classification. Conventional ontology-based algorithms [10] use the same features and train a node classifier at every ontological node to determine the node’s immediate children. However, super-categories usually possess larger intra-class variations, thus the same low-level features are often not discriminative enough to capture the common features of sub-categories. In contrast, our method learns hierarchical weak attributes for different semantic levels. The idea is more sensible and tends to follow the human recognition behavior. Furthermore, the bagging framework is used to reduce the error propagations of hierarchical classifiers. Therefore the accuracy of our method is improved.

Various feature representations have been proposed for image classification [1, 11, 6, 7, 2, 5]. Many [1, 7] are based on providing the information about the entire image. Some feature selection methods [2, 11] select important subsets of low-level features, but are limited by the semantic gaps between low-level features and high-level concepts. Other works [6] instead use manually-defined attributes, which requires both the domain knowledge and extensive labeling work. Inspired from [5], our method leverages multiple instance learning to automatically learn weak attributes, which are semantically meaningful yet no manual work needed. Our method differs from [5] in that [5] still uses the one-against-rest framework to learn weak attributes for each class independently, whereas our method learns hierarchical weak attributes for different semantic levels, and therefore additional visual cues can be obtained.

3. ONTOLOGICAL BAGGING ALGORITHM

To leverage semantic relationships for image classification, our method uses the bagging framework to train several hierarchical classifiers, each of which has the same structure as the given semantic ontology. At each ontological node, categories are grouped into super-categories based on the ontological structure and weak attributes are learned for every super-category respectively. These weak attributes are then used as image features to train a node classifier in order to discriminative between the node’s sub-categories.

In the following sections, we first describe the semantic grouping (Section 3.1). Then we elaborate on the weak attribute learning (Section 3.2) and bagging classifiers construction (Section 3.3) in details.

3.1. Semantic grouping

To construct a hierarchical classifier, we need training images of all categories in an ontology. For example, in Figure 1b, we need data of the categories “car”, “horse”, “sedan” and “hatchback”. One naive way is treating all categories independently and collecting crowdsourcing data containing all the categories. Obviously it is inefficient. A more sensible way is only collecting images of leaf categories (e.g., “horse”, “sedan” and “hatchback”). Then based on semantic relationships, we can easily obtain training images for categories at intermediate semantic levels by grouping together images of their offsprings.

Specifically, at a given ontological node m, its immediate children $m_1, ..., m_M$ are regarded as the super-categories, where M is the total number of the immediate children. All images of a particular leaf category c_i are assigned to the label m_j if the c_i is an offspring of m_j. For example, in Figure 1b, the super-categories at the root node are “car” and “horse”. Then training images of “car” will include training images of ”sedan” and ”hatchback”.

3.2. Weak attribute learning

Given images assigned with super-category labels which are obtained from Section 3.1, our algorithm automatically learns several unique weak attributes for each super-category by multiple instance learning. At each time we treat images of

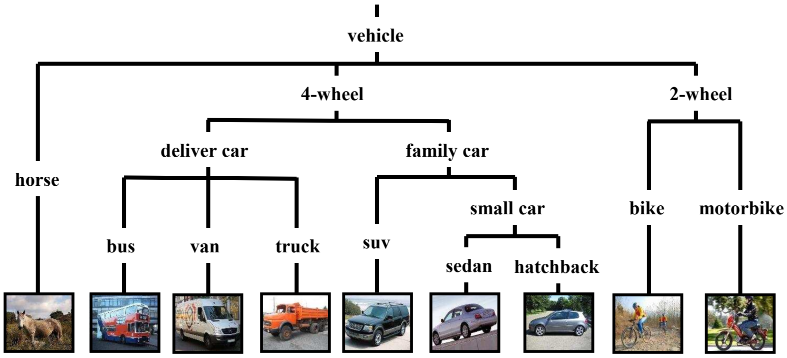


Fig. 2: The ontology for the vehicle dataset.

one particular super-category as positive bags, and the rest images as negative bags, then image windows sampled from an image (bag) will correspond to the instances of that bag.

To learn K unique weak attributes of a super-category m, we randomly select several image windows from each training image (X_i, Y_i). Every image window x_{ij} has a latent variable $z_{ij} \in \{0, 1, ..., K\}$. If $z_{ij} = k \in \{1, ..., K\}$, x_{ij} is the positive instance of the k-th weak attribute of m. Otherwise if $z_{ij} = 0$, x_{ij} is the negative instance. Weak attributes can be learned by solving the following objective function:

$$\min_{W,z_{ij}} \sum_{k=0}^{K} ||w_k||^2 + \lambda \sum_{ij} \max(0, 1 + w_{r_{ij}}^T x_{ij} - w_{z_{ij}}^T x_{ij})$$
$$s.t. \quad \text{if} \quad Y_i = m, \sum_{j} z_{ij} > 0, \text{else if} \quad Y_i \neq m, z_{ij} = 0. \tag{1}$$

where $r_{ij} = arg\,max_{k \in \{0,...,K\}, k \neq z_{ij}} w_k^T x_{ij}$. Each w_k represents the k-th positive weak attribute while w_0 denotes the negative weak attribute. Please refer to [5] for details of how to solve the objective function.

After weak attributes are learned for all the super-categories at the given node, the image feature representation can be constructed from the responses of the weak attributes. Specifically, the response of an image window x_{ij} given by the k-th weak attribute of the super-category m is $w_{mk}^T x_{ij}$. Thus, for an input image X_i, we can obtain a response map for every weak attribute. For each response map, the maximal responses are pooled over spatial pyramids. The feature concatenation of all response maps leads to the final feature descriptor for the input image.

3.3. Constructing bagging classifiers

At each internal node of the ontology, since images are assigned with their super-category labels, together with the learned feature descriptors, this allows us to learn a traditional multi-class SVM classifier. We use linear SVM classifiers for simplicity. A hierarchical classifier is constructed by collecting all the node classifiers of the ontology.

To mitigate the error propagations of a hierarchical classifier, we leverage the bagging framework to train multiple hierarchical classifiers in the same way. In addition, in order to decrease the generalization error, we insert randomness to each hierarchical classifier to make them as uncorrelated as possible. Specifically, our method randomly selects the number of weak attributes for every super-category at each node; Our method randomly selects a subset of training images to construct each hierarchical classifier. When testing, an image is classified by descending each hierarchical classifier and combining the predictions from all of them.

4. EXPERIMENTS

In this section, we evaluate our algorithm on a crowdsourcing image dataset: an object recognition dataset of 9 vehicle categories. Experimental results show that our method achieves the state-of-the-art results on this challenging dataset. We also demonstrate the advantages of our method as well as illustrate some hierarchical weak attributes for different semantic levels.

4.1. Dataset

We select 9 vehicle classes from ImageNet [8], including “horse”, “bike”, “motorbike”, “sedan”, “hatchback”, “SUV”, “van”, “truck” and “bus”. Each class has 1200 to 1700 images. We use WordNet to generate a semantic ontology for the 9 vehicle classes. We retrieve all nodes in WordNet that contains any of the class names on their word lists and build a compact hierarchical ontology by pruning the irrelevant nodes. The resulting vehicle ontology is shown in Figure 2.

4.2. Baselines

We compare our method to several baselines:

- Locality-constrained Linear Coding (LLC) [1]: uses encoded global low-level features and 3 pyramid levels to incorporate spatial information, categories are then classified by a linear SVM classifier.

Method	mean Average Precision (%)
LLC [1]	46.32 ± 1.51
SH [10]	45.61 ± 1.50
RF [3]	48.60 ± 1.26
MMDL [5]	52.27 ± 1.58
MMDL + Bagging	52.59 ± 1.19
Ours	$\mathbf{55.56 \pm 1.16}$

Table 1: Comparison of the mean average precisions (%) on the vehicle dataset. Our method outperforms all the baselines. The best result is highlighted with bold fonts.

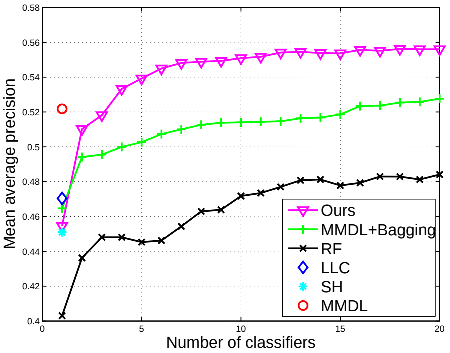


Fig. 4: Mean average precisions of different methods comparison on the crowdsourcing dataset over the number of classifiers.

- Semantic hierarchies (SH) [10]: uses the same low-level features and trains a linear SVM classifier at every ontological node.

- Random forest (RF) [3]: randomly partitions categories into a binary set at each tree node of decision trees and learns a linear SVM classifier for the splitting.

- Max-margin multiple-instance dictionary learning (M-MDL) [5]: uses the One-Against-Rest (OAR) framework to learn weak attributes for each class independently and train a multi-class linear SVM classifier for classification.

- MMDL + Bagging: leverages the bagging framework to train multiple MMDL classifiers.

4.3. Results

For each experiment run, we randomly select 30% of the training images per class and test on the remaining images. We run the experiment for 10 times and record averages and standard deviations of mean average precisions for all the algorithms. The results are shown in Table 1. Our method outperforms all the baselines.

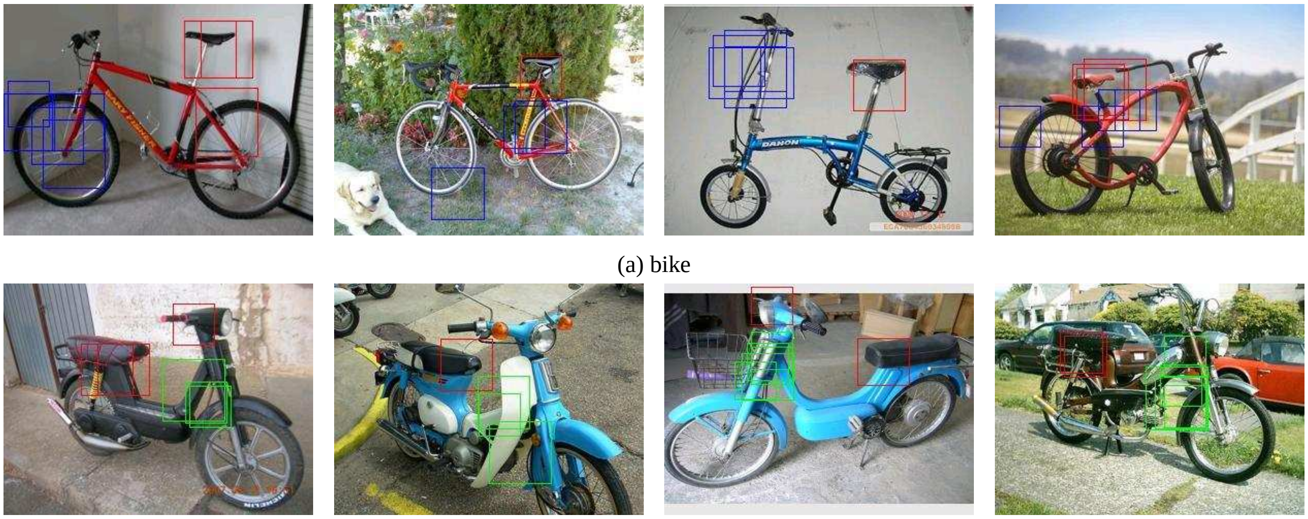
The effect of hierarchical weak attributes: One advantage of our method over the other baselines is that our method can learn weak attributes for different semantic levels. In Figure 3, we visualize some of the learned weak attributes for "bike" and "motorbike". The red bounding boxes correspond to the shared weak attribute at intermediate semantic levels while the blue and green bounding boxes relate to the unique weak attributes at leaf levels. We observe that they have strong responses on image windows which are truly semantically meaningful. Hierarchical weak attributes enable our method to classify categories from coarse to fine semantic grains by using more discriminative features, which parallels the human visual system [12]. They also describe image contents more completely, therefore achieve better classification accuracy.

The effect of bagging: From Table 1 we observe that the conventional ontological classifier SH has a lower accuracy than the OAR SVM classifier LLC, even though they use the same low-level features. This validates the conclusions from many previous works [9, 10] that the error propagations of hierarchical classifiers lead to worse classification accuracy.

Our method leverages the bagging framework to alleviate the error propagations of hierarchical classifiers. We also un-correlate the bagged classifiers in order to reduce the generalization error. The classification accuracy over the number of classifiers is shown in Figure 4. We notice that although the accuracy of using a single hierarchical classifier is low, our method can outperform all the baselines by combining the results of only 5 classifiers. This further demonstrates the effectiveness of our method. For better comparison, we also train multiple rounds of MMDL for bagging (denoted as MMDL-Bagging). However since it totally ignores semantic relationships and cannot learn hierarchical weak attributes, its result is unaffected.

5. CONCLUSION

In this work, we proposed to use ontology to incorporate semantic relationships for image classification of crowdsourced data. Our method learns discriminative features on each level of ontology using multiple instance learning, and classifies categories from coarse to fine semantic grains based on these features, which mimics the human visual system. Our method also leverages ontological knowledge to augment crowdsourcing annotations in order to construct hierarchical classifiers. Experimental results on an object recognition dataset demonstrate the effectiveness of our method. The future work is to evaluate our algorithm on other image datasets.



(a) bike

(b) motorbike

Fig. 3: The learned weak attributes for different semantic levels. Each row illustrates some examples of the category “bike” (a) or “motorbike” (b). Every bounding box indicates one positive instance of a particular weak attribute. Red bounding boxes correspond to a weak attribute of “2-wheel”. Blue bounding boxes correspond to a unique weak attribute of “bike”. Green bounding boxes correspond to a unique weak attribute of “motorbike”. (**Best viewed in color.)**

Acknowledgement

6. REFERENCES

[1] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong, “Locality-constrained linear coding for image classification,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

[2] Anna Bosch, Andrew Zisserman, and Xavier Muoz, “Image classification using random forests and ferns,” in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[3] Bangpeng Yao, Aditya Khosla, and Li 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.

[4] Zhenhua Wang, Bin Fan, and Fuchao Wu, “Local intensity order pattern for feature description,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 603–610.

[5] Xinggang Wang, Baoyuan Wang, Xiang Bai, Wenyu Liu, and Zhuowen Tu, “Max-margin multiple-instance dictionary learning,” *International Conference on Machine Learning*, 2013.

[6] Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing, “Object bank: A high-level image representation for scene classification & semantic feature sparsification,” in *Advances in neural information processing systems*, 2010, pp. 1378–1386.

[7] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 2, pp. 2169–2178.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[9] Gregory Griffin and Pietro Perona, “Learning and using taxonomies for fast visual categorization,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[10] Marcin Marszalek and Cordelia Schmid, “Semantic hierarchies for visual object recognition,” in *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*. IEEE, 2007, pp. 1–7.

[11] Kristen Grauman, Fei Sha, and Sung J Hwang, “Learning a tree of metrics with disjoint visual features,” in *Advances in Neural Information Processing Systems*, 2011, pp. 621–629.

[12] Charles A Collin and Patricia A Mcmullen, “Subordinate-level categorization relies on high s-

patial frequencies to a greater degree than basic-level categorization," *Perception & Psychophysics*, vol. 67, no. 2, pp. 354–364, 2005.