Error Recovered Hierarchical Classification

Shiai Zhu[†], Xiao-Yong Wei[‡], Chong-Wah Ngo[†]
[†]Dept of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
[‡]College of Computer Science, Sichuan University, Chengdu, China
shiaizhu2@student.cityu.edu.hk, cswei@scu.edu.cn, cwngo@cs.cityu.edu.hk

ABSTRACT

Hierarchical classification (HC) is a popular and efficient way for detecting the semantic concepts from the images.

However, the conventional HC, which always selects the branch with the highest classification response to go on, has the risk of propagating serious errors from higher levels of the hierarchy to the lower levels. We argue that the highest-response-first strategy is too arbitrary, because the candidate nodes are considered individually which ignores the semantic relationship among them. In this paper, we propose a novel method for HC, which is able to utilize the semantic relationship among candidate nodes and their children to recover the responses of unreliable classifiers of the candidate nodes, with the hope of providing the branch selection a more globally valid and semantically consistent view. The experimental results show that the proposed method outperforms the conventional HC methods and achieves a satisfactory balance between the accuracy and efficiency.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

Concept Detection, Large-scale Hierarchy, Error Propagation

1. INTRODUCTION

Detecting semantic concept(s) from images is a multi-class classification problem, which has received intensive studies in the last decade. The complexity of the problem increases dramatically when the number of concepts exceeds a certain extent. To reduce the complexity, a practical way is to simplify the problem into a set of binary classification problems, where a binary classifier is trained for each concept, and during testing, all classifiers are applied to the target to which class label(s) are assigned by the classifier(s) with the highest response(s). While the *one (target)-vs-all (concepts)* simplification provides a much more feasible solution for multiple concept detection, nevertheless, it is still impractical when facing large-scale datasets (e.g., ImageNet [4] which includes 21,841 concepts with each of them associated with 1,000 images), because all classifiers have to be called at runtime for every image.

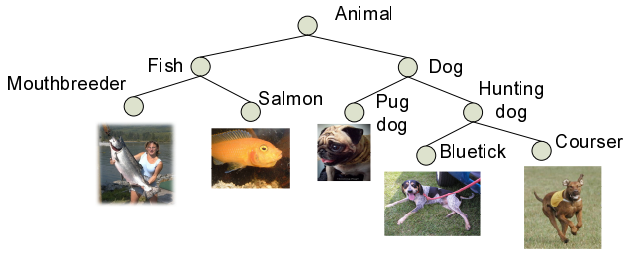

Figure 1: An example of image concept hierarchy.

To tackle the scalability issue, hierarchical classification is commonly adopted. Instead of applying all classifiers to a target image, hierarchical classification organizes concept classifiers into a hierarchy (e.g., Figure 1) according to the semantic relationship among concepts, and only selects a small set of classifiers for testing. The selection procedure usually starts from the root of the hierarchy and proceeds with a top-down manner that, for each node under investigation, hierarchical classification first applies the classifiers of the child nodes to the target, and then selects the child node with the highest response on its classifier as the next node to investigate. The procedure repeats recursively and results in a path from the root to a leaf node (e.g., Animal-Fish-Salmon), on the basis of which all the concept labels on the path will be assigned to the target image.

By reducing the number of classifiers to be visited, hierarchical classification significantly improves the efficiency of multiple concept detection, and thus has been widely employed (e.g, web categorization [3, 1] and gene function prediction [10]). However, as pointed out by Bennett et al [1], the improvement is paid at the price of scarifying effectiveness of the classification. More specifically, the top-down classification procedure will make the classification errors included at the higher levels of the hierarchy be propagated to the lower levels, and in turn significantly degrades the accuracy of those leaf nodes. A few methods have been proposed to address this problem. In [3], Xue et al. address the problem by selecting only a small number of nodes from the original hierarchy to construct a simplified hierarchy for

*Xiao-Yong Wei is the corresponding author.

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502182.

classification. In this way, the chance of error propagation is reduced because the classification path from root to leaf node is significantly shorten. In [1], when building the set of negative instances for training each node, Bennett et al. also include the false positive instances which have been misclassified at its ancestor nodes, in the hope that those instances can be rejected by current classifier and the mis-classification errors will not be further propagated to the low level nodes.

Even those methods are able to improve the accuracy of hierarchical classification, we argue that the optimization schemes employed still have not addressed the core of the error propagation – the arbitrariness of the branch selection. In the top-down selection procedure, the classification will be led to a branch under the classifier with the highest response. This is arbitrary because it is well known that the performance of the visual concept detectors is still not satisfactory in a general sense, and therefore, it frequently happens that the high response is from an unreliable classifier. Selecting a branch under such node may seriously ruin the classification procedure follows. In addition, the highest-response-first strategy is lack of global perspective, in the way that the classification only focuses on the local responses of the nodes to be investigated, but never verifies if these responses are valid or globally consistent to the semantic relationship encapsulated in the hierarchy. With Figure 1 as an example, assuming that a target image includes a fish, the classification will be led to the branch under Dog if the classifier for Dog is unreliable and outputs higher response than that of Fish. When deciding next node to move on, the classifiers of the Pug Dog and Hunting Dog (if reliable enough) will output low responses. Globally speaking, this is conflicting to the semantic relationship in the hierarchy, in the sense that a parent node is with high response while all of its children nodes are with low responses. However, the procedure will never "doubt" the decision at the node Dog and continue to select between Pug Dog and Hunting Dog, resulting in further error propagation.

In this paper, we propose a novel branch selection scheme for hierarchical classification to address the core issue related to error propagation. Instead of arbitrarily selecting the node with the highest response, we introduce an error recovery scheme which first verifies the semantic consistency of the observation on a candidate node under investigation and those of its relatives (e.g., siblings, children and so on) in the hierarchy, and then adjusts the output of the node accordingly. The decision of which branch to go will be delayed when the verifications of all candidate nodes are finished. In other words, the proposed scheme makes the decision only when more observations are available and it is "confident" to do so. In the example mentioned above, the scheme is able to detect the inconsistency among the observation of Dog and those of its children nodes, and thus increase the chance of leading the classification to the correct branch (i.e., that under Fish in this example).

2. METHOD

In the proposed method, instead of considering only the candidate nodes, we also involve their children and siblings to form a committee for decision making. In practical, before deciding which branch to go, we adjust the response of each candidate node to be semantically consistent with those of other nodes in the committee, with the hope that the unreliable response of a candidate can be fixed if it is conflicting with those of its relatives in the hierarchy. Therefore, making decision on the adjusted response is with a more global view and can avoid the arbitrariness of the branch selection with the highest-response-first local strategy.

2.1 Problem Formulation

Given an instance x and a node c_t as the current node in a hierarchy, we denote the branch to go at next moment as a node c_{t+1}, so that

$$c_{t+1} = \arg\max_{c \in \mathcal{N}(c_t)} \hat{f}_c(x) \tag{1}$$

where $\mathcal{N}(c_t)$ is a set of child nodes of c_t, and $\hat{f}_c(x)$ is the adjusted response of c to x. Further denoting the original response of c as $f_c(x)$, we can formulate the highest-response-first local strategy by replacing $\hat{f}_c(x)$ with the original response $f_c(x)$. Moreover, our committee for decision making (denoted as $\mathcal{T}$) is a union of c_t's children (i.e., $\mathcal{N}(c_t)$) and its grandchildren (i.e., $\bigcup_{c \in \mathcal{N}(c_t)} \mathcal{N}(c)$). The problem to solve is then how to define the adjusting function $\hat{f}_c(x)$ with respect to both the semantical relationship and the observations of the nodes in the committee $\mathcal{T}$. Let us denote the semantical relationship among nodes as $\Phi_{\mathcal{T}}$ and compose the observations of the committee $\mathcal{T}$ into a vector $\mathbf{f}_{\mathcal{T}}(x) = [f_{c_1}, f_{c_2}, f_{c_3}, \cdots]$ where $c_1, c_2, c_3, \cdots \in \mathcal{T}$, the problem can be formulated as

$$\hat{f}_c(x) = P(c|\Phi_{\mathcal{T}}, \mathbf{f}_{\mathcal{T}}(x)). \tag{2}$$

2.2 Committee-based Response Adjustment

Once the hierarchy is known, there are a lot of priori can be utilized for modeling $\hat{f}_c(x)$. For example, by definition, the siblings in a hierarchy are semantically exclusive, so that the response for a candidate node should be approaching 1 if those of its siblings are all with responses close to 0. In addition, parent node represents a union of the instances of its child nodes, so that the response for a candidate node should be close to 0 if those of its child nodes are all with responses close to 0. In brief, confined by the semantic relationship $\Phi_{\mathcal{T}}$, the responses of nodes in a committee should always follow certain patterns. Therefore, we can use the observations of the committee $\mathcal{T}$ to predict that of a candidate node so as to implement the response adjustment.

Intuitively, this can be simply modeled by logistic regression, where we use the observations of the committee $\mathcal{T}$ as predictors for estimating a reasonable output for a target concept c. The semantic constraints $\Phi_{\mathcal{T}}$ is then modeled by a set of weights (i.e., a weight vector $\mathbf{w}_c = [w_1, w_2, \cdots]$) associated with the predictors. A weight given to a predictor reflects the ability of the predictor to estimate the output of the target concept. By further expanding the logistic regression to all the candidate nodes, we can learn their weights at the same time by multi-class regression (MLR), resulting a weight matrix $\mathbf{W}$. It is worth mentioning that learning the weights together not only brings convenience for the learning but also makes the inter-concept relationship among candidate nodes be modeled during the learning. By replacing the semantic relationship $\Phi_{\mathcal{T}}$ with $\mathbf{W}$, Eq. (2) can be implemented with MLR as

$$P(c|\mathbf{W}, \mathbf{f}_{\mathcal{T}}(x)) = \frac{\exp(\mathbf{w}_c^T \mathbf{f}_{\mathcal{T}}(x))}{\sum_{c_k \in \mathcal{N}(c_t)} \exp(\mathbf{w}_k^T \mathbf{f}_{\mathcal{T}}(x))}, \tag{3}$$

Table 1: Dataset statistics: number of leaf nodes (#Leaf) and internal nodes (#Int), depth of the hierarchy (#Dep), average number of instances of each concept for training (#Trn), validation (#Val) and testing (#Tst) respectively.

Dataset	#Leaf	#Int	#Dep	#Trn	#Val	#Tst
Caltech256	256	62	6	58	29	29
ILSRVC1K	1000	645	13	1261	50	150

where $\mathbf{w}_k$ is the weight vector for the corresponding candidate node $c_k \in \mathcal{N}(c_t)$.

Given a set of training instances $\mathbf{X} = \{x_1, x_2, \cdots\}$ with each of them associated with a class label $y_i \in \mathcal{N}(c_t)$, an optimal weight matrix $\mathbf{W}^*$ can be obtained by

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} - \sum_{x_i \in \mathbf{X}}^{|\mathbf{X}|} \log P(\mathbf{y}_i|\mathbf{W}, \mathbf{f}_{\mathcal{T}}(x_i)) + \lambda \|\mathbf{W}\|^2, \quad (4)$$

where the second term is a regularizer used to control the model complexity, and λ is regularization parameter. Equ. 4 is referred to as L2-regularized MLR. This problem can be efficiently solved by Quasi-Newton method. In the experiment, we adopt the package released by Mark Schmidt*.

2.3 Verification and Recovery

Since the labels given to each instance always follow certain patterns which reflect the inter-concept semantic relationship in the hierarchy, the resulting weight matrix **W*** in Equ. 4 is also embedded with those relationship which can be used to verify if a set of responses is consistent to those patterns, in the way that it results in larger response in Eq. (3) when it is consistent and smaller response otherwise. Note that, during learning, we put the candidate node itself in the committee, with the hope that the resulting weight can also reflect the classification reliability of the candidate node. According to the principle of MLR, if a candidate node is with an unreliable classifier, it will be assigned with a small weight to weaken its impact to the final results (i.e., the adjusted response Eq. (3)), and the predication for its label will mainly rely on the responses of other nodes in the committee. By contrast, if the node is a reliable classifier, it earns a large weight so its impacts will dominate those of others. This also explains why we put the candidate node itself in the committee. Therefore, the proposed method fulfills the semantic relationship verification and error recovery at the same time by MLR.

3. EXPERIMENTS

3.1 Settings

We use two image datasets for evaluation, Caltech256 [2] and ILSVRC1K [4]. Caltech256 consists of 256 labeled concepts for object annotation. In [2], a concept hierarchy is pre-defined using the 256 concepts as leaf nodes. ILSRVC1K is a subset of ImageNet, where the concepts are organized by WordNet. Starting from the 1,000 concepts in ILSRVC1K, a hierarchy is extracted from the ImageNet hierarchy. Table 1 lists the statistics. We follow the train/validation/test split in ILSRVC1K. For Caltech256, the instances of each concept are split to train/val/test by 50%-25%-25%.

*http://www.di.ens.fr/ mschmidt/Software/code.html

Table 2: Performance comparison of three hierarchical classification methods on Caltech256 and ILSRVC1K. Classification performance is measured by global accuracy (*Acc*), and testing efficiency is measured by average saved time cost (*MTC*).

Dataset		HC	SIB-HC	ER-HC
Caltech256	*Acc* (%)	26.7	27.0	30.5
	MTC (%)	91.5	91.1	67.1
ILSRVC1K	*Acc* (%)	9.4	9.6	11.2
	MTC (%)	98.1	98.3	88.7

Table 3: Performance comparison of three hierarchical classification methods on ILSRVC1K. Local classifiers are learnt on free-sampled web images.

	HC	SIB-HC	ER-HC
Acc (%)	4.3	5.1	5.6
MTC (%)	98.2	97.9	89.4

Each image is represented using Locality-constrained Linear Coding (LLC) with densely sampled SIFT features [6]. We employ a visual vocabulary of 4,000 visual words, and three level spatial partitions (1×1, 2×2 and 3×1). Consequently, the dimension of feature vector is 32,000. For each node c_i, a classifier $f_{c_i}(x)$ is learnt using linear SVM [7] on training set. In addition, a multi-class logistic regression model is learnt for each internal node on the validation set.

The testing instances are all from leaf nodes which are contextually exclusive. Thus we evaluate the classification performance using global accuracy among leaf nodes (*Acc*). An instance is correctly classified when the annotated leaf node hits the ground-truth. Similar to [5] where the efficiency of HC is measured using one-vs-all approach as baseline, we evaluate the efficiency using percentage of saved time cost compared to one-vs-all approach. Since we adopt linear SVM, the time cost is linear with the number of involved classifiers. Thus the saved time cost is defined as $TC = 1 - \frac{\#model}{\#concept}$, where $\#model$ and $\#concept$ denote the number of activated classifiers and the number of concepts in the hierarchy respectively. In this case, TC of one-vs-all approach is 0. We further define MTC as the average saved cost over all testing instances.

We use the standard hierarchical classification (HC) that employs the highest-response-first strategy as our baseline. Our proposed error recovered HC is referred to as ER-HC. To investigate the impact of different committees, we implement a simplified ER-HC by using only the candidate nodes to form the committee (i.e., only sibling relationship is considered). We denote this method as SIB-HC.

3.2 Performance on Benchmark Datasets

The results on two datasets are summarized in Table 2. For Caltech256, ER-HC improves the baseline by 14%, while by only taking the sibling relationship into account, SIB-HC only improves the baseline slightly by 1%. This result demonstrates the advantage of a larger committee which postpones the decision making until more observations are available and it is more confident to do so. In terms of efficiency, ER-HC sacrifices more computational cost than those of HC and SIB-HC, but the saved cost by 67.1% from the one-vs-all approach is still considered significant, an indication that ER-HC achieves a better balance between com-

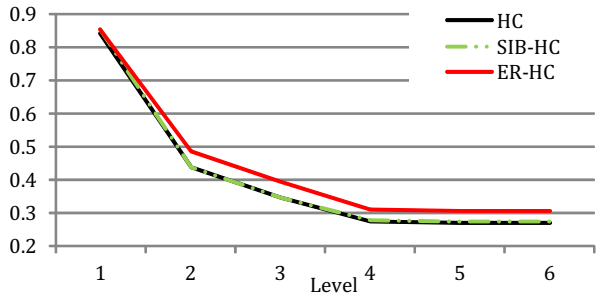

Figure 2: Accuracy on Caltech256 by level

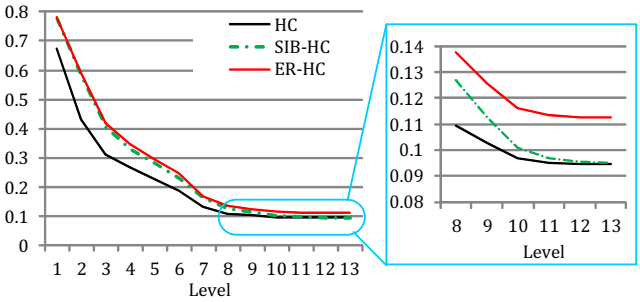

Figure 3: Left: accuracy on ILSRVC1K by level. Right: zoom-in from level 8 to level 13.

putational cost and accuracy. Similarly, for ILSRVC1K, ER-HC outperforms HC and SIB-HC by 19% and 16.6% in accuracy respectively, while remaining satisfactory in efficiency by saving 88.9% computational cost from the baseline. We can see that the advantage of ER-HC on ILSRVC1K is more obvious than that on Caltech256. This is due to the fact that the hierarchy of ILSRVC1K is in a larger scale which includes 1,327 more concepts than Caltech256. Therefore, the SVM models for nodes at higher levels of the ILSRVC1K hierarchy are with much more complex or vague decision boundaries than those in Caltech256, because more offspring concepts are attached to each high level node which bring the instances with more variances of visual appearances, and finally make the resulting classifier less reliable. In this case, the semantic relationship verification and error recovery embedded in MLR model of ER-HC are more necessary, because the error propagation is more serious.

To grasp more insights of the methods, we plot the accuracy of each method at each hierarchy level on the two datasets in Figure 2 and Figure 3 respectively. The error propagation becomes more serious with the increase level of depth, resulting in drop of accuracy. ER-HC has demonstrated consistent superiority over HC and SIB-HC, confirming its ability to address the issue of error propagation. Surprisingly, the performance of SIB-HC seems is better on ILSRVC1K than on Caltech256. This again confirms our analysis that the advantage of employing a committee for decision making over the arbitrary highest-response-first strategy will be more obvious when the classifiers of individual nodes are weaker.

3.3 Cross-dataset Evaluation

We also study the feasibility of applying the proposed approach on classifiers learnt from non-expert labeled examples. These classifiers are usually "weaker" classifiers and suffer from the problem of domain shift [9]. For this purpose, we adopt the approach in [8] to automatically crawl at most 1,000 positive training instances from Web for each leaf category. As a result, there are around 950,000 images being downloaded to train 1,645 classifiers. The results are shown in Table 3. The advantage of the proposed approach is more evidenced on classifiers that are learnt from these noisy training samples. Compared to that in Table 2, the accuracy of all methods degrades a lot. This is not surprising because free-sampled Web images are noisy and the domain-shift from training dataset to testing is more significant, which make most of the classifiers unreliable. However, ER-HC exhibits the best performance in accuracy with 27% improvement over baseline. The proposed approach thus has potential to be scalable to classifiers that are trained using non-expert labels, which is an important feature for large-scale multimedia applications as annotating training examples is a highly expensive process especially when thousands of classifiers are in demand.

4. CONCLUSION

We have presented the error recovered hierarchical classification approach, which utilizes the semantic relationship among concept nodes in a hierarchy for addressing the error propagation problem. The method has demonstrated significant and consistent performance improvements over conventional methods on Caltech256, ILSRVC1K, and free-sampled Web image dataset, while maintaining a satisfactory balance between efficiency and accuracy. Currently, we only consider the single label hierarchy where leaf nodes are contextually exclusive. Future work includes the extension of current work for multi-label hierarchical classification. We will also address the issue on reducing the error propagation, for example, by traversing the tree starting from reliable internal nodes, rather than from the root node.

5. ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 61272290, Grant 61228205 and Grant 61001148.

6. REFERENCES

[1] P. N. Bennett and N. Nguyen. Refined experts: Improving classification in large taxonomies. In *SIGIR*, 2009.
[2] G. Griffin et al. The caltech-256. *Caltech Technical Report*, 2007.
[3] G.-R. Xue et al. Deep classification in large-scale text hierarchies. In *SIGIR*, 2008.
[4] J. Deng et al. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
[5] J. Deng et al. Fast and balanced: Efficient label tree learning for large scale object recognition. In *NIPS*, 2011.
[6] J. Wang et al. Locality-constrained linear coding for image classification. In *CVPR*, 2010.
[7] R. E. Fan et al. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.
[8] S. A. Zhu et al. Sampling and ontologically pooling web images for visual concept learning. *IEEE Trans. on MM*, 14(4):1068–1078, 2012.
[9] T. Yao et al. Predicting domain adaptivity: Redo or recycle? In *ACM MM*, 2012.
[10] Z. Barutcuoglu et al. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.