Reducing Error Propagation for Hierarchical Classification

Tien-Dung Mai[a,*], Thanh Duc Ngo[a], Duy-Dinh Le[a,b],
Duc Anh Duong[a], Kiem Hoang[a], Shin'ichi Satoh[b]

[a] *University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam*
[b] *National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan*

Abstract

Hierarchical classification is an efficient approach in terms of prediction accuracy at a fraction of the prediction cost . One of the main challenging problems of this approach is to deal with error propagation. Irrelevant branching decision made at a parent node cannot be corrected at its child nodes in traversing the tree for classification. This leads to decrease classification accuracy. In this paper we introduce a novel approach to reduce branching error at a node by exploiting its relative relationship. Given a parent node on the tree, we develop a new model that allows representing each candidate branch basing on classification response of its child nodes, grandchild nodes and their differences with siblings. We then formulate the selecting candidate branch as a maximum-margin problem. By optimizing this problem, we obtain a classifier that can determinate the most discriminating candidate. Empirical results demonstrate that our proposed approach outperforms related approaches on benchmark datasets such as Caltech-256, SUN-397 and ILSVRC2010-1K.

Keywords: Hierarchical classification, error propagation, node relationship, large-scale image classification

*Corresponding author
Email addresses: dungmt@uit.edu.vn (Tien-Dung Mai), thanhnd@uit.edu.vn (Thanh Duc Ngo), ledduy@nii.ac.jp (Duy-Dinh Le), ducda@uit.edu.vn (Duc Anh Duong), kiemhv@uit.edu.vn (Kiem Hoang), satoh@nii.ac.jp (Shin'ichi Satoh)

1. Introduction

Multi-class image classification is a challenges problem with a large number of classes, dimensions and images. A widely used approach is one-versus-all strategy where an independent classifier is learnt per class. However, the classifying cost might be expensive since all the classifiers would need to be evaluated every time for predicting a given test image. This is impractical when the number of classes is large [1], [2], [3], [4].

Recently, several hierarchical classification approaches (i.e. tree-based approaches) [5], [6],[7], [8], [9], [10], [11] have been proposed to tackle that scalability issue. The key idea of these approaches is to organize a given set of classes into a hierarchical tree structure in which each node is associated with a subset of class labels and a classifier. The classifier produces a response indicating how relevant a new image or test sample to the class labels of the node. A leaf node associates with only one class. A root node is associated with the entire set of classes. Classifying a given test sample is achieved by traversing the tree from its root until it reaches a leaf node. At each visited node, classifiers of their children are applied. The returned scores from the classifiers are then used to determinate the next node to move on. With such top-down traversing procedure, tree-based approaches have a well-known problem - propagation error problem. Errors made at a higher level in the label tree cannot be corrected at later levels [8], [9]. The origination of this error is caused by tree organization. In tree-based model, at each node v of the tree that it has at most Q children per node for branching, a set of class labels $\ell(v)$ of node v are partitioned into Q groups, each group corresponds to a child node, and classifiers of nodes are then learned. Due to the number of class labels $|\ell(v)|$ is often greater than the number of children Q, it is difficult to obtain Q groups clearly discriminated. Moreover, in the tree building process from training set, to guarantee balanced tree structure for efficiency, an equal number of class labels in each group (or associated with each child node) should be maintained. This leads to highly related classes can be partitioned into different groups.

As a result, the highest response score of these classifiers is not confident enough to make the correct decision every time. So, only selecting the child node with the highest response score to move on (Fig. 1(a)) - referred as the highest-response-first strategy - might decrease classification accuracy. .

To tackle above mentioned problems, the relation between scores with others has been recently considered. Liu et al. [8] introduced an method that allows visiting nodes with highest score and the second highest score if the difference between them is below a threshold value. Although this way is able to improve accuracy, it is not only increasing the classification cost when selecting occur at higher level node, but also difficult to determinate a adaptive global threshold in practical. Zhu et al. [9] proposed an method that concatenates the scores of candidate nodes and of their children into a meta feature vector that is used to select visited node by a multi-class regression. However, this method ignores information about their relative difference.

In this work, our objective is to further extend the effectiveness of using relationship of relative nodes that is more accurate at branching decision than the highest-response-first based approach and the approach of Zhu et al. [9]. Our main contributions are follows:

- First, we do not only use response scores from classifiers of child nodes and grandchild nodes of the current node. We further extend by also considering their relative differences. Given a parent node on the tree, we model each candidate branch by considering classification response of its child nodes, grandchild nodes and their differences with siblings at different levels (Fig. 1(c)). All of such information is composed into a feature vector representing the candidate branch.

- Second, we formulate branching decision making as an optimization problem. By solving the problem, we obtain a maximum margin classifier by which the most discriminating candidate branch can be selected. Additionally, our formulation is closely related to Multiple Instance Learning (MIL). Thus, it can take the advantages of MIL to deal with weakly la-

beled data in training process.

We evaluate the approach on several benchmark including Caltech-256, SUN-397 and ILSVRC2010-1K. Experimental results indicated that our proposed approach improve classification accuracy, compared to other related approaches.

The rest of the paper is organized as follows. In Section 2, related works are presented. In Section 3, the proposed method is described. The experimental results are presented in Section 4. Finally, Section 5 concludes the paper.

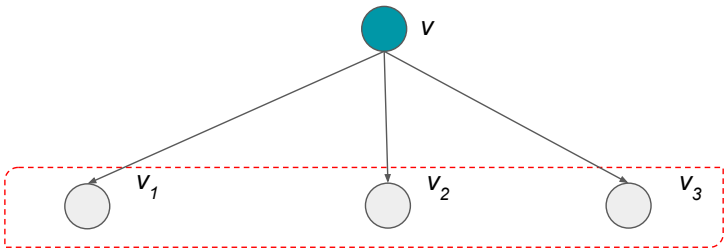
2. Related work

A challenge problem in a tree-based approaches is how to use the tree for accurate classification. The popular method usually selects the branch or node whose corresponding predict score is highest to follow in the next step. In cases where there is ambiguity in the correct node, this selection will reduce the classification accuracy if the wrong node is selected.

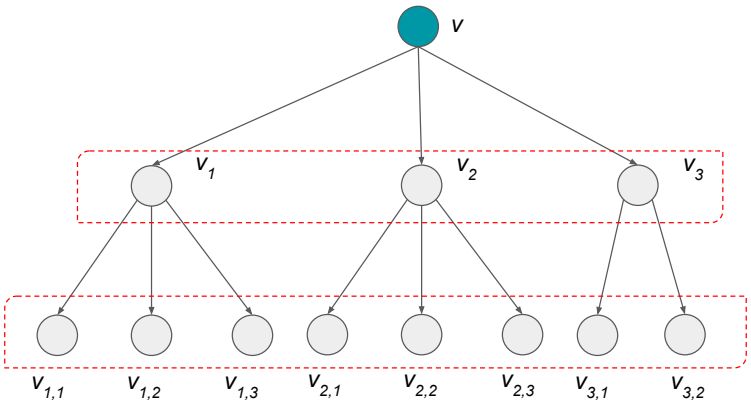
Providing such label trees, avoiding irrelevant branching decision is essential to improve classification accuracy. Beside the highest-response-first strategy, several other approaches introduced recently.

Sun et al. [10] considered the classification problem as finding the best path in the label tree and proposed a branch-and-bound-like algorithm to efficiently search for the best path. To trade-off between efficiency and accuracy, both the bounds and the classifiers are jointly learned using structured SVM formulation with additional bound constraints.

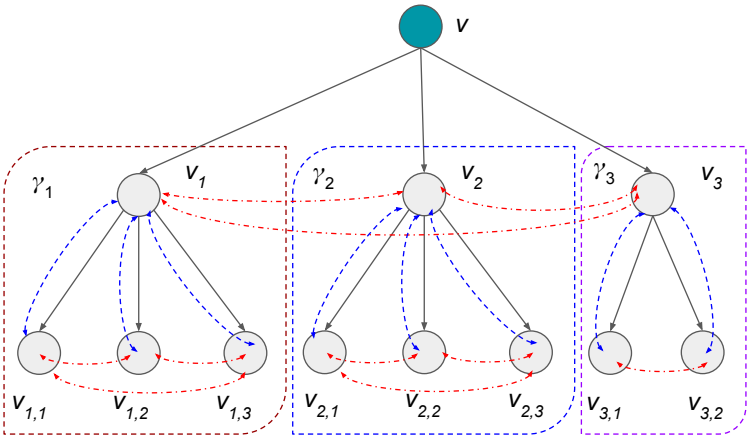
Wang and Forsyth in [7] proposed a method that aggregate the probability distribution associated with leaf node of trees in a label forest i.e. an ensemble of label trees. The $(i+1)$-th label tree is constructed by applying Bengio et al. method [5] with confusion matrix computed for the i-th label tree on a validation set. Although this method improved the classification accuracy, the computational cost can be significantly increased with a large number of label trees.



(a) Presenting the baseline approach which relies on highest-response-first strategy



(b) illustrates Zhu et al.'s method [9] relying on a meta feature constructed by concatenating the prediction scores of child nodes and grand child nodes.



(c) Illustrating our proposal. The approach takes into account the relationships between child nodes, grand child nodes and their siblings.

Figure 1: Different approaches to select the next node from the current node v

Liu et al. [8] introduced a method that rises the classification accuracy by performing a limited search over multiple branches of the label tree. A choice is made by comparing the difference in scores between the highest-scoring and the second highest-scoring branch. If this difference is below a threshold, both branches are traversed. Although, the classification accuracy is increased as the classification process explores more branches, the computational efficiency is decreased and the value of threshold is manually tuned to balance the accuracy and efficiency trade-off. However, this indicates that the relationship among nodes can help to improve classification accuracy.

In [9], Zhu et al. has found that the classification error may be reduced by using the semantic and contextual relationship among candidate nodes and their children to recover the responses of unreliable classifiers of the candidate nodes. They simply concatenated the scores of candidate nodes and of their children into a meta feature vector that is used to predict candidate node by using multi-class regression (Fig. 1(b)).

3. Our Approach

In contrast with prior research, we investigate a novel technique to perform effective selection which best branch to follow. In this section, we first describe a method to represent the relationship between candidate nodes and of its children as sets of feature vectors in subsection 3.1 , and then the node selection is formulated as an optimize problem in subsection 3.2, and finally the classification for new testing images is described in subsection 3.3

3.1. Representing Node Relationships

In this work, we take into account relationships between child nodes and grandchild nodes together with their siblings, instead of simply concatenating response scores from classifiers at nodes as in [9]. A candidate branch from the current node is represented by using: i) the prediction score from classifier of its child node along the branch; ii) the prediction score from classifier of its

grandchild node along the branch; ii) differences in prediction scores between the child node and the grandchild node with other sibling nodes at the same level.

Specifically, let $p_v(x)$ be a prediction score of image x returned by the classifier of node v, and $\ell(v)$ denotes the set of class labels that belong to node v. $\psi(v)$ is the set of children of node v, $v_i \in \psi(v)$ indicates the i-th child of node v and $v_{i,j} \in \psi(v_i)$ indicates the j-th child of node v_i.

Suppose that we are considering the current node v, the task is to select one of its child nodes to visit next. Noted that we select one child node here for simplification. In general, more than one node can be selected. Let $\rho_{v_i}(x) \in \mathbb{R}^{|\psi(v)|}$ is a vector composed by the prediction score $p_{v_i}(x)$ and the difference between $p_{v_i}(x)$ and those of siblings of v_i. It can be represented as follows:

$$\rho_{v_i}(x) = [p_{v_i}(x), p_{v_i}(x) - p_{v_1}(x), .., p_{v_i}(x) - p_{v_{|\psi(v)|}}(x)] \quad (1)$$

Then, a feature vector that describes the relation between a node v with its child node, its grandchild nodes and differences to their siblings can be formulated as:

$$\varphi_{i,j}(x) = [\rho_{v_i}(x), \quad \rho_{v_{i,j}}(x)] \quad (2)$$

Noted that $\varphi_{i,j}(x)$ also represents a candidate branch which traverses through v_i and $v_{i,j}$.

3.2. Problem Formulation

For each node v_i, we have the set of candidate branches go through v_i is:

$$\gamma_{v_i}(x) = \{\varphi_{i,1}(x), \varphi_{i,2}(x), .., \varphi_{i,|\psi(v_i)|}(x)\} \quad (3)$$

Accordingly, the set of all possible branches from v is $\gamma_{v_i}(x), \forall v_{i,j} \in \psi(v_i)$. Then, the problem of selecting the next node to move on becomes selecting suitable v_i. We formulate the problem as a max margin optimization problem.

Assuming we are given an image x with label y. A candidate branch represented by a feature vector $\varphi_{i,j}(x)$ is labeled positive if the label $y \in \ell(v_i)$ and $y \in \ell(v_{i,j})$. Otherwise, the branch is labeled negative.

A set $\gamma_{v_i}(x)$ representing all possible branches go through v_i is labeled positive if it contains at least one positive branch. It is labeled negative if all of its branches are negative. Selecting a candidate node v_i to follow is equivalent to finding a corresponding positive set $\gamma_{v_i}(x)$.

Using images having ground-truth class labels belong to $\ell(v)$ of a node v, we can obtain positive sets and negative sets. We use Γ_v^+ to denote the set of all positive sets $\gamma_{v_i}(.)$ and Γ_v^- to denote the set of all negative sets $\gamma_{v_i}(.)$.

We then find a margin function with parameters (w, b) to maximize the margin between positive set Γ_v^+ and negative set Γ_v^- by solving the following optimization problem:

$$\arg \min_{w_v, b, \{\xi_i\}} \frac{1}{2}\|w_v\|^2 + C \sum_i \xi_i \tag{4}$$

subject to

$$\max_{\gamma \in \Gamma_v^+} (w_v^T \cdot \gamma + b) \geq 1 - \xi_i, \quad \forall i = 1, ..., |\Gamma_v^+|, \xi_i \geq 0 \tag{5}$$

$$\max_{\gamma \in \Gamma_v^-} (w_v^T \cdot \gamma + b) \leq -1 + \xi_j, \quad \forall j = 1, ..., |\Gamma_v^-|, \xi_j \geq 0 \tag{6}$$

where C is a constant, $\{\xi_i\}$ are non-negative slack variables.

This problem is solved as in [12], we obtain a classifier that can be used to determinate a candidate node v_i to follow in the next step.

3.3. Classification Stage

Given a new testing image x, classification are performed by traversing the tree from its root to a leaf node. At each node v along the path, we first obtain a set Γ_v^x. Γ_v^x includes $|\psi(v)|$ sets of candidate branches go through $v_i, v_i \in \psi(v)$.

$$\Gamma_v^x = \{\gamma_{v_1}(x), .., \gamma_{v_{|\psi(v)|}}(x)\} \tag{7}$$

We then select node v_i whose $\tilde{\gamma}$ yielding the largest.

$$\tilde{\gamma} = \arg\max_{\gamma \in \Gamma_v^x} (w_v^T \cdot \gamma + b) \tag{8}$$

We summarize the algorithm for classification in Algorithm 1 as follows.

Algorithm 1 Classification

Input: The label tree $T = (V, E)$ with the root node r and a testing image x

Output: Class label of x.

1: $v \leftarrow r$
2: **while** $\sigma(v) \neq \emptyset$ **do**
3: Obtain Γ_v^x by equation (7)
4: $v \leftarrow v_i$ correspond to $\tilde{\gamma}$ obtained by equation (8).
5: **end while**
6: **return** $l(v)$

4. Experiments

4.1. Datasets

We conducted experiments on several benchmark which are widely used to evaluate large-scale image classification approaches.

- **Caltech-256** dataset [13]. This is a dataset for multi-class object classification. There are 29,780 images of 256 object classes. Each image is assigned to a single class label, each class contains at least 80 images.
- **SUN-397** dataset [14]. This dataset is a subset of the SUN dataset. It is selected from 908 scene classes for a scene classification. There are 108,754 images of 397 classes with at least 100 images per class.
- **ILSVRC2010-1K** dataset [3]. This is a challenging dataset, especially for large-scale image classification. The dataset includes 1,461,406 images of 1,000 classes. Each class has at least 668 images.

The dataset is divided into a training set, a validation set and a test set. The training set is used to learn classifier at each non-leaf node of tree. The validation set is used to train maximum margin classifier (proposed approach), multi-class regression (described in [9]), and to obtain confusion matrix used for building tree structure (described in [5]).

In the case of ILSVRC2010-1K, we use the provided validation set and test set. We randomly pick 300 images per category for training.

With Caltech-256 and SUN-397, the images of each category are randomly split to 50% for training, 25% for validation and remaining images for testing. To obtain stable experimental results, the random split was performed 5 times. And then, we reported the average classification performance with the corresponding standard deviation.

Feature: We follow the widely used feature settings as in [6], [8]. In particular, each image is represented using Locality-constrained Linear Coding (LLC) described in [15] with densely sampled SIFT features extracted by VLFeat toolbox [16]. A codebook with 10,000 visual words is used together with two-level Spatial Pyramid Matching (SPM) [17] with 1×1 and 2×2 grids. Consequently, the number of dimensions of each feature vector is 50,000.

4.2. Experimental Setting

4.2.1. The Tree Configures

In order to show the performance of our proposed approach, we conducted experiments with three following configurations:

- The pre-defined tree. This is a tree structure that created manually basing on semantic relations or the hierarchical structure of WordNet.

- The label tree. We employ the approach of Bengio et al. [5] for tree construction. Specifically, binary one-versus-all classifiers firstly are trained form training set. These classifiers are applied on a validation set to obtain a confusion matrix. A spectral clustering algorithm [18] is then applied recursively to partition the label sets of a current node into disjoint groups

of classes. Each group associates to one child node of the current node. Since spectral clustering penalizes unbalanced partitions, the resulting tree structure can be unbalanced. The tree is denoted by T_Q and it has at most Q children per node for branching.

- The balanced label tree. We consider a case of balanced label tree [19] to achieve the logarithmic run-time. The tree structure is built by jointly optimizing the balance constraint and the confusion constraint during learning. The balanced tree is denoted by $T_{Q,H}$ that it has at most Q children per node for branching and the maximum height is H.

4.2.2. Baseline approaches

We compare the performance of proposed approach with to closely related approaches.

- Baseline. This is the popular approach which relies on the highest-response-first strategy for selecting a next branch.

- ER-SHC [9]. We directly compare to this approach since it is the most related approach to our proposal. Both approaches aim at improving classification accuracy by employing relationships of nodes. This direct comparison is to realize the effectiveness of utilizing more information of node relationships into branch selection and max margin formulation.

4.3. Results and Discussions

4.3.1. Results on The Predefined Tree

Table 1 shows the experimental results on the pre-defined tree of ILSVRC2010-1K, SUN-397, and Caltech-256, respectively. Each column of table is associated with dataset, and each row shows the performance of methods. As we can see, the performance of our proposed method is better than others.

4.3.2. Results on The Label Tree

The performance of the evaluated approaches on ILSVRC2010-1K is reported in Table 2. For all tree configurations, our proposed approach achieved

Table 1: Performances of methods on predefined tree using SIFT-LLC-SPM feature.

Methods	ImageNet-1K	SUN-397	Caltech- 256
Standard	x ±x	37.39 ±0.45	35.45 ±0.34
ER-SHC [9]	x ±x	39.57 ±0.48	36.43 ±0.34
Ours	x ±x	41.34 ±0.32	36.85 ±0.28

significant accuracy improvement over other approaches. The approach of Zhu et al. performed slightly better than the baseline approach on this dataset. A consistent observation is that accuracy of all approaches drops as the maximum number of branches Q of the tree T_Q is reduced. This can be explained by the fact that the smaller Q is the larger number of classes are partitioned into the same group (or branch). It might cause groups more confused; thus, makes precise prediction more challenging.

Furthermore, as Q is decreased, the tree becomes higher. It results in higher chance of choosing an irrelevant branch when traversing. Besides, the accuracy gap between relationship based approaches (i.e. ours and ER-SHC) and the baseline approach is narrowed as Q decreases, since we have less information from nodes for branching decision.

Experimental results on Caltech-256 and SUN-397 are presented in 5 and Table 4 respectively. On SUN-397, our approach improve 2% accuracy over the approach by Zhu et al. [9] and 3% accuracy over the Baseline approach. Similarly, our approach outperformed other approaches on Caltech-256, except with T_2. With this tree, the feature vector representing each candidate branch is generated with very limited amount of information from nodes. As result, the max margin classifier cannot produce precise selection.

4.3.3. Results on The Balanced Tree

The classification performance for different balanced tree configures on ImageNet-1K, SUN-379, and Caltech-256 are presented Table 3, Table 6, and Table 7, respectively. It can be seen that the our proposed results are better performance

Table 2: Performance of the evaluated approaches on ILSVRC2010-1K with different tree configurations. The tree is built by Bengio et al.'s method [5] using SIFT-LLC-SPM feature.

Methods	T_{32}	T_{10}	T_6	T_4
Baseline	7.32 ±0	6.01 ±0	5.52 ±0	5.12 ±0
ER-SHC [9]	7.70 ±0	5.70 ±0	5.12 ±0	4.66 ±0
Ours	**12.68 ±0**	**8.48 ±0**	**6.76 ±0**	**6.04 ±0**

Table 3: Performances of methods on ILSVRC2010-1K on tree configurations built by Mai et al.'s method [19] using SIFT-LLC-SPM feature.

Methods	$T_{32,2}$	$T_{10,3}$	$T_{6,4}$	$T_{4,5}$
Baseline [19]	14.07 ±0	12.89 ±0	12.61 ±0	12.20 ±0
ER-SHC [9]	15.95 ±0	13.37 ±0	12.27 ±0	11.97 ±0
Ours	**20.91 ±0**	**16.13 ±0**	**14.36 ±0**	**13.96 ±0**

Table 4: Performance of the evaluated approaches on SUN-397 with different tree configurations. The tree is built by Bengio et al.'s method [5] using SIFT-LLC-SPM feature.

Methods	T_{20}	T_8	T_5	T_4
Baseline	30.42 ±0.79	24.07 ±0.96	23.11 ±0.72	21.33 ±0.34
ER-SHC [9]	30.63 ±0.78	24.35 ±1.36	23.73 ±0.62	21.77 ±0.42
Ours	**34.39** ±0.57	**28.21** ±1.21	**25.95** ±0.60	**23.71** ±0.28

Table 5: Performance of the evaluated approaches on Caltech-256 with different tree configurations. The tree is built by Bengio et al.'s method [5] using SIFT-LLC-SPM feature.

Methods	T_{16}	T_7	T_4	T_2
Baseline	31.55 ±0.43	27.48 ±0.52	24.64 ±0.24	22.60 ±0.48
ER-SHC [9]	31.15 ±0.77	27.43 ±0.31	24.95 ±0.23	**22.63** ±0.56
Ours	**37.08** ±0.27	**29.32** ±0.52	**26.45** ±0.48	22.26 ±0.15

Table 6: Performances of methods on SUN-397 on tree configurations built by Mai et al.'s method [19] using SIFT-LLC-SPM feature.

Methods	$T_{20,2}$	$T_{8,3}$	$T_{5,4}$	$T_{4,5}$
Baseline [19]	38.49 ±0.75	35.94 ±0.89	33.90 ±0.52	32.94 ±0.70
ER-SHC [9]	40.72 ±0.01	37.05 ±0.11	34.99 ±0.44	33.87 ±0.12
Ours	**43.04 ±0.01**	**38.86 ±0.11**	**35.76 ±0.44**	**34.05 ±0.12**

Table 7: Performances of methods on Caltech-256 on tree configurations built by Mai et al.'s method [19] using SIFT-LLC-SPM feature.

Methods	$T_{16,2}$	$T_{7,3}$	$T_{4,4}$	$T_{2,8}$
Baseline [19]	38.95 ±0.43	35.22 ±0.52	33.18 ±0.24	29.15 ±0.48
ER-SHC [9]	39.10 ±0.48	35.49 ±0.68	33.42 ±0.60	**29.21 ±0.46**
Ours	**43.53 ±0.46**	**36.54 ±0.20**	**34.42 ±0.39**	28.50 ±0.37

than others.

4.3.4. Experiments on Deep Feature

We carried out additional experiments on Caltech-256 using the the state-of-the-art deep visual VGG-VERYDEEP-16 feature [20]. Each image is represented by a feature vector with 4,096 dimensions.

Experimental results with the balanced tree [19] are showed in Table 8. As you seen, that the accuracy of the proposed method is higher than that of the others.

4.3.5. Discussions

- Compexity of proposed approach: higher than Baseline, same with [9]. Accuracy is higher than.
- Effecting of Q and Accuracy (Fig.2)

Table 8: Performances of methods on Caltech-256 on tree configurations built by Mai et al.'s method [19] using VGG-VERYDEEP-16 feature.

Methods	$T_{16,2}$	$T_{7,3}$	$T_{4,4}$	$T_{2,8}$
Baseline [19]	72.83 ±0.45	69.60 ±0.65	68.38 ±0.25	**65.89 ±0.54**
ER-SHC [9]	73.77 ±0.32	71.03 ±0.48	68.97 ±0.24	65.83 ±0.37
Ours	**75.18 ±1.09**	**71.42 ±0.75**	**69.05 ±0.28**	64.94 ±0.54
OvA	79.23 ± 0.42			

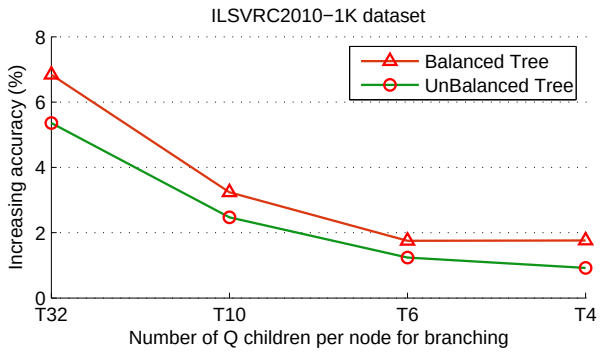


Figure 2: Effecting of Q and Accuracy on ILSVRC2010-1K dataset

5. Conclusion and Future Work

In this paper, we introduced an approach to reduce propagation error in hierarchical classification by utilizing relationships of nodes. First, given a parent node on the tree, we model each candidate branch by considering classification response of its child nodes, grandchild nodes and their differences with siblings at different levels. Second, we formulate branching decision making as an optimization problem. Experiments on several benchmarks indicated that the proposal help to improve classification accuracy, compared to related approaches.

References

[1] J. Deng, A. C. Berg, K. Li, L. Fei-Fei, What does classifying more than 10, 000 image categories tell us?, in: ECCV, 2010, pp. 71–84.

[2] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Good practice in large-scale learning for image classification, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2014) 507–520.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV) (2015) 1–42doi:10.1007/s11263-015-0816-y.

[4] H. O. Song, R. B. Girshick, S. Zickler, C. Geyer, P. F. Felzenszwalb, T. Darrell, Generalized sparselet models for real-time multiclass object recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (5) (2015) 1001–1012.

[5] S. Bengio, J. Weston, D. Grangier, Label embedding trees for large multi-class tasks, in: Advances in Neural Information Processing Systems 23, NIPS 2010 . Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada., 2010, pp. 163–171.

[6] J. Deng, S. Satheesh, A. C. Berg, F. Li, Fast and balanced: Efficient label tree learning for large scale object recognition, in: Advances in Neural

Information Processing Systems 24, NIPS 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain., 2011, pp. 567–575.

[7] Y. Wang, D. A. Forsyth, Large multi-class image categorization with ensembles of label trees, in: Proceedings of the 2013 IEEE International Conference on Multimedia and Expo, ICME 2013, San Jose, CA, USA, July 15-19, 2013, 2013, pp. 1–6.

[8] B. Liu, F. Sadeghi, M. F. Tappen, O. Shamir, C. Liu, Probabilistic label trees for efficient large scale image classification, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013, 2013, pp. 843–850.

[9] S. Zhu, X. Wei, C. Ngo, Collaborative error reduction for hierarchical classification, Computer Vision and Image Understanding 124 (2014) 79–90.

[10] M. Sun, W. Huang, S. Savarese, Find the best path: An efficient and accurate classifier for image hierarchies, in: IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013, 2013, pp. 265–272.

[11] R. Babbar, I. Partalas, E. Gaussier, M. reza Amini, On flat versus hierarchical classification in large-scale taxonomies, in: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (Eds.), Advances in Neural Information Processing Systems 26, 2013, pp. 1824–1832.
URL http://media.nips.cc/nipsbooks/nipspapers/paper_files/nips26/926.pdf

[12] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, in: Advances in Neural Information Processing Systems 15, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada], 2002, pp. 561–568.

[13] G. Griffin, A. Holub, P. Perona, Caltech-256 object category dataset, Tech.

Rep. 7694, California Institute of Technology (2007).
URL http://authors.library.caltech.edu/7694

[14] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, A. Torralba, SUN database: Large-scale scene recognition from abbey to zoo, in: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, 2010, pp. 3485–3492.

[15] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, 2010, pp. 3360–3367.

[16] A. Vedaldi, B. Fulkerson, Vlfeat: an open and portable library of computer vision algorithms (2010).

[17] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA, 2006, pp. 2169–2178.

[18] A. Y. Ng, M. I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: Advances in Neural Information Processing Systems 14, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada], 2001, pp. 849–856.

[19] T. Mai, T. D. Ngo, D. Le, D. A. Duong, K. Hoang, S. Satoh, Learning balanced trees for large scale image classification, in: ICIAP, 2015, pp. 3–13.

[20] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, in: British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5, 2014, 2014.