MMF: Multi-Task Multi-Structure Fusion for Hierarchical Image Classification⋆

Xiaoni Li[1,2], Yucan Zhou[1,✉], Yu Zhou[1], and Weiping Wang[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{lixiaoni, zhouyucan, zhouyu, wangweiping}@iie.ac.cn

Abstract. Hierarchical classification is significant for complex tasks by providing multi-granular predictions and encouraging better mistakes. As the label structure decides its performance, many existing approaches attempt to construct an excellent label structure for promoting the classification results. In this paper, we consider that different label structures provide a variety of prior knowledge for category recognition, thus fusing them is helpful to achieve better hierarchical classification results. Furthermore, we propose a multi-task multi-structure fusion model to integrate different label structures. It contains two kinds of branches: one is the traditional classification branch to classify the common subclasses, the other is responsible for identifying the heterogeneous superclasses defined by different label structures. Besides the effect of multiple label structures, we also explore the architecture of the deep model for better hierachical classification and adjust the hierarchical evaluation metrics for multiple label structures. Experimental results on CIFAR100 and Car196 show that our method obtains significantly better results than using a flat classifier or a hierarchical classifier with any single label structure.
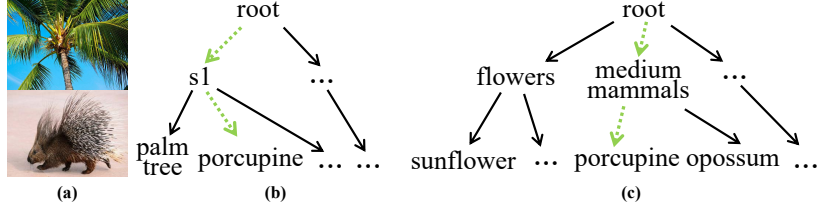
Keywords: Hierarchical classification · Multi-task learning · Multiple label structures.

1 Introduction

Although deep learning in text spotting [27,28,6,5,26,25], object detection [37], self-supervised learning [40,39,23,38,22] and image classification [10,15] has achieved dramatic performance with the increase of annotated data, the unclassifiable categories are growing and inevitable in the ear of big data. Moreover, the conventional one-hot coding in flat classifiers suggests a strict error evaluation: as long

Fig. 1. The benefit of combining multiple label structures. (a) shows two samples for palm tree and porcupine, as these two categories are similar, they are easy to be misclassified. (b) and (c) are two label structures, each for the visual sturcture based on the affinity matrix and the semantic structure. The second layers in (b) and (c) are the superclasses, the third layers are the shared subclasses. The green dashed paths are the ground truth in both label structures when a "porcupine" is needed to identify. When "s1"and "medium mammals"are recognized in each label structure, "porcupine"is promoted as it belongs to both "s1"and "medium mammals".

as the predicted value is inconsistent with the real one, it will be recognized as misclassification. In fact, there are different levels of severity in mistakes [4]. As shown in Figure 1(c), the classifier makes a less serious mistake when it classifies a "porcupine" into an "opossum" than a " sunflower" obviously, because they all belong to the superclass "medium mammals". Therefore, when misclassification is unavoidable, providing a reasonable mistake is more significant.

Recently, more and more work is devoted to using hierarchical classification methods [35,34] to make multi-granular predictions and avoid serious mistakes. In hierarchical classification, label structures play a critical role. Hence many researchers try to construct efficient label structures, which can be roughly divided into semantics-based methods and computation-based methods. The former extracts the semantic structure from WordNet [18], where categories are organized into a tree-shape structure according to their semantic relations [41,7,8,14]. However, these relations may be inconsistent with the appearances, which weakens the performance of classification tasks. Therefore, a lot of work builds visual information tree structures [13,3,19,20,21,16,11,29]. Some build the tree structures based on the confusion matrix [13,3,19,20,21], which is constructed by the results of a classifier. Others construct the label structure based on the affinity matrix [16,11,29] calculted by the similarity of any two categories.

Different label structures provide various prior knowledge for the underlying classification tasks. Hence integrating these structures can further improve the performance [33,43]. As shown in Fig.1, in the mission of "porcupine" classification, if one has determined its superclass "s1" and "medium mammals" according to the label structure based on the affinity matrix and semantics respectively, then, "porcupine" can be easily determined by combining these two intermediate results. A straightforward strategy to fuse multiple label structures is constructing a hierarchical classifier for each structure, and the prediction is obtained by integrating the results of multiple classifiers [33]. This idea is sim-

ple and efficient, but in the deep learning scenario, it is memory-consuming and computationally redundant to design a neural network for each label structure.

In this paper, a multi-task multi-structure fusion (MMF) model is proposed to make the superclasses from different label structures instruct the subclass recognition. It achieves this by encouraging the learned feature to satisfy the multiple similarity constraints in various hierarchical label structures. Specifically, it is a deep convolutional neural network with two kinds of classification branches: the conventional classification branch (CCB) used for identifying subclasses, and the multiple superclass classification branches (MSCBs), where each branch is responsible for recognizing the superclasses defined by a specific label structure.

Our main contributions are summarized in three folds: 1) We find that integrating multiple label structures can further improve the performance of hierarchical classification, and propose a MMF model to combine different hierarchical label structures. 2) Further, various architectures of our MMF model are explored for better classification. 3) We adjust the hierarchical evaluation metrics for multiple label structures. Experimental results on CIFAR100 and Car196 are better than traditional flat classifiers and hierarchical classifiers with any single label structure.

2   Related Work

2.1   Hierarchical Classification

The traditional methods decompose the hierarchical classification task into several subtasks and train a subclass classifier for each superclass node independently [13,3,29,9]. However, this strategy is memory-consuming and computing expensive for storing and training many subclass classifiers. Therefore, these methods are not suitable for deep learning. For deep hierarchical classification, Frome *et al.* [12] constructs a deep visual-semantic model by re-training the lower layers of the pre-trained visual network to predict the vector representation of the image label text in the hierarchical label structure learned by the language model. Barz & Denzler [2] design an algorithm to map the labels into a unit hypersphere where the cosine distances between different labels are equal to the distance in the hierarchical label structure.

Besides the implicit label embedding, many researchers want to explicitly model the hierarchical label structure. Wu *et al.* [36] adds one fully connected softmax layer for each layer in the hierarchy to make the network recognize both the superclasses and the subclasses. But in this work, the relations between the superclasses and subclasses are underutilized. Bertinetto *et al.* [4] adds a weight matrix between the superclass classifier and the subclass classifier, thus, predictions of the superclass can be propagated to and affect the predictions of the subclass through the weight matrix. Ahmed *et al.* [1] trains a network to provide superclasses information and common knowledge through shared features to a set of expert networks, each of which devoted to recognizing the subclasses of a specific superclass. Therefore, the multi-task framework has been proved efficient for hierarchical classification.
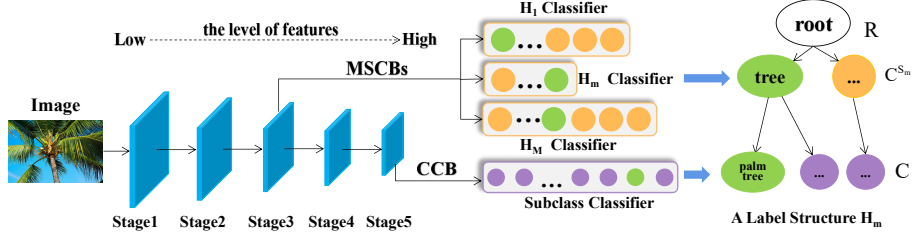
Fig. 2. The architecture of our MMF model with a five-stage CNN.

2.2   Multiple Label Structures Fusion

As we have mentioned, different label structures provide different priori knowledge for hierarchical classification, thus, integrating these structures can further improve the performance. Wang *et al.* [33] constructs a classifier for each label structure, and the prediction of a test sample is obtained by integrating the results of multiple classifiers. This idea is simple and efficient, but in the deep learning scenario, it is memory-consuming and computationally redundant to design a neural network for each label structure. Instead of training multiple subclass classifiers, Zhao *et al.* [43] fuses multiple category similarities defined by different label structures in the kernel space, then trains one kernel SVM classifier. Inspired by this idea, we propose a multi-task multi-structure framework to make the superclasses from different label structures instruct the subclass recognition by encouraging the learned features to satisfy the multiple similarity constraints in different label structures.
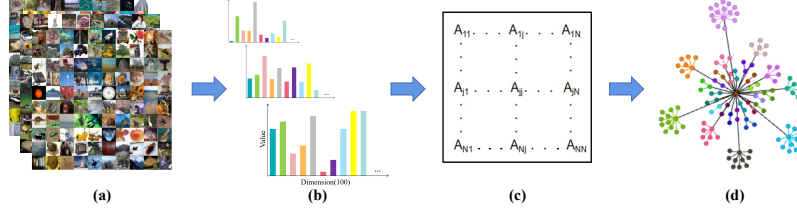
3   Method

3.1   Problem Definition

Given an image dataset $\mathbf{D}$ with N classes, after M label structure construction methods applied, we can obtain M tree-like label structures. Except for the layer containing the root node, each layer in the structure is equipped with a specific classifier to decide the category in the current layer. To simplify the problem, all the structures covered in this paper are arranged with three levels. Take the right side of Fig.2 as an example, a label structure is represented as $\mathbf{H_m} = \{\mathbf{R}, \mathbf{C^{S_m}}, \mathbf{C}\}$, where $\mathbf{R}$ is the root node, $\mathbf{C^{S_m}}$ is the superclass set in $\mathbf{H_m}$, and $\mathbf{C}$ is the subclass set. Consequently, given a sample x in $\mathbf{D}$, its labels compose of one subclass c and M superclasses c^{s_m}. For hierarchical classification with multiple label structures, all these superclasses should be predicted.

3.2   Multi-Task Multi-Structure Fusion Framework

Assumed that multiple label structures have been obtained, the MMF model can be constructed, which includes two kinds of classification branches: CCB is a classification branch with a traditional classifier to identify the subclass, while

Fig. 3. The construction of $\mathbf{H_A}$ on CIFAR100. (a) Samples from CIFAR100 dataset. (b) Feature representations devided from 100 classes in images. (c) The affinity matrix obtained from the features. (d) $\mathbf{H_A}$ with 30 superclasses after spectral clustering.

MSCBs are several classifiers for the superclass identification. Fig.2 shows an overview of our model. All MSCBs work in parallel to encourage the features derived from the network to meet various similarity constraints in different label structures, guiding CCB to make more accurate predictions.

MSCBs The MSCBs contains multiple superclass classifiers. As shown in Fig.2, the "$\mathbf{H_m}$ classifier" in MSCBs completes its task based on the label structure $\mathbf{H_m}$. As the superclasses are more generic than the subclasses (e.g., medium mammals and porcupine in Fig.1 (c)), and the high-level features usually contain more details to discriminate the subclasses, MSCBs should be inserted in the early stages. We will explore the influence of various network stages to attach the MSCBs in the following experiments.

3.3   Multiple Label Structures

For a dataset $\mathbf{D}$, we introduce two kinds of $\mathbf{H_m}$ to construct our MMF model: the semantic label structure $\mathbf{H_S}$ and the visual label structure $\mathbf{H_A}$ based on the affinity matrix.

$\mathbf{H_S}$ Usually, a semantic structure is adopted to organize the data. Take CIFAR100 as an example: there is a three-level semantic hierarchical label structure (one root node, 20 superclasses, and 100 subclasses) in the dataset (like $\mathbf{H_m}$ in Fig.2). We adopt this semantic structure $\mathbf{H_S}$ inherent in the datasets in our paper.

$\mathbf{H_A}$ We construct a visual label structure based on the affinity matrix through two stages as shown in Fig.3: feature extraction and label structure construction. For feature extraction (from (a) to (b)), we use a pre-trained VGG16 to extract features. Then, for label structure construction, we adopt sample pairwise distance to calculate the similarity between any two categories (from (b) to (c)), and simplify the calculation with [29] by Eq.(1), where c_i is the i-th class, and Q_{c_i}, σ_{c_i} are the mean and variance of features in c_i. Then the affinity matrix $\mathbf{A}$ can be constructed by Eq.(2), where δ_{ij} is a self-tuning parameter [30], and we take 1 in our work. Finally, we use spectral clustering [24] to build the corresponding $\mathbf{H_A}$ (from (c) to (d)).

$$dis(c_i, c_j)^2 = \left\| Q_{c_i} - Q_{c_j} \right\|^2 + \sigma_{c_i}^2 + \sigma_{c_j}^2. \tag{1}$$

$$A_{ij} = \exp(-\frac{dis(c_i, c_j)}{\delta_{ij}}). \tag{2}$$

3.4   Hierarchical Measures

As it is a multi-structure fusion work, we add the hierarchical information to the evaluation measures, and consider the similarity between the predicted class and the ground truth, *i.e.*, the severity of the classifier's mistakes. However, the existing evaluation measures [42] such as hierarchical F_1-measure (F_H), the tree induced loss (TIE) and the lowest common ancestor (LCA) of the prediction and the ground truth, are designed for a single label structure. Therefore, we adjust the above three measures to fit our method.

F_{Ha} The traditional precision P and recall R rate are extended to the hierarchical precision P_H and recall R_H rate, which can well measure the severity of mistakes, as the error in the superclasses is more serious than that in the subclasses. As our MMF deals with multiple label structures, we take the average of all P_H and R_H in each label structure. F_{Ha} is calculated from P_{Ha} and R_{Ha}:

$$P_{Ha} = \frac{1}{M} \sum_{m=1}^{M} \frac{\left| C_{aug}^{\hat{m}} \cap C_{aug}^{m} \right|}{\left| C_{aug}^{\hat{m}} \right|}, \ R_{Ha} = \frac{1}{M} \sum_{m=1}^{M} \frac{\left| C_{aug}^{\hat{m}} \cap C_{aug}^{m} \right|}{\left| C_{aug}^{m} \right|}, \tag{3}$$

$$F_{Ha} = \frac{2 \cdot P_{Ha} \cdot R_{Ha}}{P_{Ha} + R_{Ha}}, \tag{4}$$

where M is the number of the label structures, $C_{aug}^{\hat{m}}$ is the predicted extension set which contains the class nodes on the path from the root class to the predict subclass in $\mathbf{H_m}$, C_{aug}^{m} is the real extension set which contains the class nodes on the path from the root class to the real subclass in $\mathbf{H_m}$, and $|\cdot|$ is an operator to calculate the number of the elements.

TIE_a In the tree structure, the total number of edges from the predicted node to the real node along a specific label structure is represented as TIE distance. To deal with multiple label structures, we introduce TIE_a to average all the TIE distances in each label structure by Eq.(5), where $|Edge_m(c, \hat{c})|$ is the number of edges from the predicted node $\hat{c}$ to the real node c in $\mathbf{H_m}$. Accordingly, the smaller the TIE_a, the more similar the predicted class is to the real class.

$$TIE_a = \frac{1}{M} \sum_{m=1}^{M} |Edge_m(c, \hat{c})|. \tag{5}$$

LCA_a We modified the LCA height to the mean value of all LCA heights in each label structure to obtain LCA_a by Eq.(6), where $Height_m(c, \hat{c})$ is the lowest common ancestor height between the predicted node $\hat{c}$ and the real node c in $\mathbf{H_m}$. A smaller LCA_a means a smaller classification error.

$$LCA_a = \frac{1}{M} \sum_{m=1}^{M} Height_m(c, \hat{c}). \tag{6}$$

3.5   Traing and Inference

The multi-task loss for MMF model contains a CCB loss and several MSCBs losses denoted by Eq.(7), where $\phi(x; \theta)$ is a classification network, the parameter θ is learned by minimizing our loss function. $\hat{c}$, $c^{\hat{s}_m}$ are the predicted subclass and superclass, c, c^{s_m} are the ground truth of subclass and superclass in $\mathbf{H_m}$ respectively. λ_m is the constraint intensity of "$\mathbf{H_m}$ classifier", and $\lambda = \sum_{m=1}^{M} \lambda_m$. We use the standard cross entropy loss to compute L_{CCB} and L_{H_m}.

$$\mathcal{L}(\phi(x, \theta), c, \mathbf{C^S}) = (1 - \lambda) * L_{CCB}(\hat{c}, c) + \sum_{m=1}^{M} \lambda_m * L_{H_m}(c^{\hat{s}_m}, c^{s_m}). \quad (7)$$

When training, samples with different hierarchical label structures are input into the framework for multiple rounds of iterative training. MSCBs impose constraints on the network through the multi-task loss, affecting the prediction of subclasses. When it comes to inference, the final predicted result of subclass is decided by CCB only.

4   Experiments

4.1   Experimental Settings

Datasets We conduct experiments on two benchmark datasets CIFAR100 and Car196. In CIFAR100, there is a total number of 100 categories belonging to 20 semantic superclasses on average. Car196 is a fine-grained dataset containing 196 subclasses from three different kinds of semantic superclasses "Make" (49 categories), "Type"(18 categories), and "Year". We choose "Make" and "Type" as the semantic label structures because "Year" is not discriminative. We also construct a three-level $\mathbf{H_A}$ for each dataset.

    Backbones The backbones of our network are VGG16 [31] and ResNet50 [32] trained from scratch. Note that there are five stages in both backbones.

    Evaluation Metrics Four evaluation metrics are considered in our work to fully analyze the classifiers' results. Besides the flat measure top-1 accuracy (Acc), we also adopt three hierarchical measures proposed before to better evaluate the performance of the classifiers.

4.2   Ablation Study

The impact of $\mathbf{H_A}$, the network stages to attach MSCBs, and the constraint intensity λ on the model's performance is explored in the following ablation experiments. Note that all the ablation studies are adapted both VGG16 and ResNet50 backbone on CIFAR100 and Car196, in order to show the generalization of our model.

$\mathbf{H_A}$s   As $\mathbf{H_A}$ is three-level, the number of superclasses decides its structure. To obtain a suitable number of superclasses, we perform a series of ablation experiments on our MMF model with a singe label structure $\mathbf{H_A}$. Referring to

Table 1. The subclass performance with different $\mathbf{H_A}$.

Dataset	Num_{super}	VGG16				ResNet50			
		Acc(↑)	F_{Ha}(↑)	TIE_a(↓)	LCA_a(↓)	Acc(↑)	F_{Ha}(↑)	TIE_a(↓)	LCA_a(↓)
CIFAR100	18	72.67	**84.15**	**0.9509**	**0.4754**	79.20	**88.12**	**0.7130**	**0.3565**
	20	72.51	84.07	0.9558	0.4779	79.01	88.04	0.7176	0.3588
	25	72.46	83.78	0.9733	0.4867	79.13	87.93	0.7240	0.3620
	30	**72.95**	84.04	0.9574	0.4787	**79.21**	87.91	0.7252	0.3626
Car196	15	**82.67**	**91.70**	**0.4979**	**0.2490**	**90.19**	**95.44**	**0.2738**	**0.1369**
	18	81.04	90.87	0.5478	0.2739	89.17	95.06	0.2963	0.1482
	20	82.42	91.47	0.5116	0.2558	89.69	95.22	0.2865	0.1433
	30	79.58	89.38	0.6370	0.3185	89.06	94.47	0.3321	0.1660
	40	79.09	88.66	0.6804	0.3402	89.26	94.33	0.3402	0.1701
	50	79.76	89.05	0.6569	0.3285	89.67	94.50	0.3297	0.1649

the number of superclasses in $\mathbf{H_S}$, We vary the number of $\mathbf{H_A}$'s superclasses in [18, 20, 25, 30] for CIFAR100 , and [15, 18, 20, 30, 40, 50] for Car196. According to the results of Table 1, we select the $\mathbf{H_A}$ with 30 superclasses for CIFAR100, and 15 for Car196.

Fig. 4. (a) MSCBs attached in different stages on CIFAR100. (b) and (c) are the constraint intensities of MSCBs on CIFAR100 and Car196 respectively. Note that results in the first row are for VGG16, and the second are for ResNet50.

MSCB$_s$ An important thing for our MMF model is where to insert the classifiers for superclasses. We explore it with $\mathbf{H_A}$, $\mathbf{H_S}$ and multiple structures $\mathbf{H_{A\&S}}$ on CIFAR100, and the results with $\lambda = 0.2$ are shown in Fig.4(a). One interesting phenomenon can be observed in the both backbones: adding the superclass classifiers in the early stages is more effective. The reason may be that the low-level features are more generic and lose details of the high-level features for subclasses identification. So in the experiments, we insert MSCBs in the early stages to make our MMF model firstly grasp general concepts, then the CCB captures details in each concept to discriminate subclasses.

Table 2. The subclass performance of VGG16.

Dataset	Method	Structure	Num_{class}	λ	Acc($\uparrow$)	F_{Ha}($\uparrow$)	TIE_a($\downarrow$)	LCA_a($\downarrow$)
CIFAR100	Greedy [3,17]	$\mathbf{H_S}$	100/20	-	70.09	81.98	1.0811	0.5405
	NBPath [29]	$\mathbf{H_S}$	100/20	-	70.49	82.23	1.0664	0.5332
	MMF	w/o $\mathbf{H}$	100	-	72.20	83.35	0.9991	0.4995
		$\mathbf{H_A}$	100/30	0.1	73.27	**84.97**	**0.9015**	**0.4507**
		$\mathbf{H_S}$	100/20	0.1	73.25	84.70	0.9179	0.4590
		$\mathbf{H_{A\&S}}$	100/30/20	0.15	**73.37**	84.79	0.9127	0.4563
Car196	Greedy [3,17]	$\mathbf{H_T}$	196/18	-	51.45	73.89	1.5663	0.7832
		$\mathbf{H_M}$	196/49	-	54.05	75.40	1.4763	0.7381
	NBPath [29]	$\mathbf{H_T}$	196/18	-	52.99	74.71	1.5172	0.7586
		$\mathbf{H_M}$	196/49	-	55.30	76.05	1.4373	0.7186
	MMF	w/o $\mathbf{H}$	196	-	74.63	87.88	0.7273	0.3637
		$\mathbf{H_A}$	196/15	0.6	82.67	91.70	0.4979	0.2490
		$\mathbf{H_T}$	196/18	0.4	82.35	91.34	0.5344	0.2672
		$\mathbf{H_M}$	196/49	0.3	82.24	91.89	0.4864	0.2432
		$\mathbf{H_{A\&T}}$	196/15/18	0.3	82.67	92.23	0.4661	0.2330
		$\mathbf{H_{A\&M}}$	196/15/49	0.3	81.62	91.77	0.4938	0.2469
		$\mathbf{H_{T\&M}}$	196/18/49	0.2	80.69	91.44	0.5134	0.2567
		$\mathbf{H_{A\&T\&M}}$	196/15/18/49	0.2	**83.67**	**92.88**	**0.4274**	**0.2137**

λ: We fix MSCBs on the stage where the best performance is achieved, then vary λ in [0.1, 0.8]. In Fig.4(b) and (c), experimental results on the subclasses show that different Acc obtained by adjusting λ. With a larger λ ($\lambda \geq 0.1$), the performance on the subclasses is worse than the MMF w/o $\mathbf{H}$. And it's not weird that results of different label structures don't coincide exactly because they have different similarity constraints, corresponding to different constraint strengths. In the following experiments with a single label structure, we set λ with the best performance. And for multiple label structures, λ_m is set to the same values for the sake of making these label structures act equally, varying within the range of the λ which achieved the best results in the single label structures.

4.3   Experimental Results and Analyses

Our deep MMF model with different single label structures and their combinations is compared with two methods based on the top-down strategy. For the top-down methods, we choose two methods which are based on the greedy selection at each hierarchy (Greedy) [3,17] and the N-Best Path (NBPath) [29]. To improve the performance, we adopt features extracted from a carefully fine-tuned VGG16 or ResNet50, which is the backbone in our MMF model. Then kernel SVMs are employed as the classifiers at each hierarchy. For our MMF model, we adopt different label structures as shown in Table 2 and Tabel 3. "w/o $\mathbf{H}$" means MMF model without any hierarchical structures, which is a traditional classification network contains a backbone and a classifier for subclass classification. "$\mathbf{H_T}$" and "$\mathbf{H_M}$" are the semantic structures based on "Type" and "Make" respectively, and "$\mathbf{H_{A\&S}}$" (i.e., $\mathbf{H_A}$ and $\mathbf{H_S}$) *etc.* are multiple label structures.

Table 2 and Table 3 show the results of VGG16 and ResNet50 on CIFAR100 and Car196, respectively. Note that in the multi-structure models, the perfor-

Table 3. The subclass performance of ResNet50.

Dataset	Method	Structure	Num_{class}	λ	Acc($\uparrow$)	F_{Ha}($\uparrow$)	TIE_a($\downarrow$)	LCA_a($\downarrow$)
CIFAR100	Greedy [3,17]	$\mathbf{H_S}$	100/20	-	76.22	85.68	0.8594	0.4297
	NBPath [29]	$\mathbf{H_S}$	100/20	-	76.44	85.78	0.8535	0.4267
	MMF	w/o $\mathbf{H}$	100	-	78.50	87.22	0.7668	0.3834
		$\mathbf{H_A}$	100/30	0.1	79.38	88.14	0.7114	0.3557
		$\mathbf{H_S}$	100/20	0.1	79.51	88.28	0.7034	0.3517
		$\mathbf{H_{A\&S}}$	100/30/20	0.15	**79.52**	**88.28**	**0.7029**	**0.3514**
Car196	Greedy [3,17]	$\mathbf{H_T}$	196/18	-	86.80	93.30	0.4022	0.2011
		$\mathbf{H_M}$	196/49	-	87.30	93.57	0.3855	0.1928
	NBPath [29]	$\mathbf{H_T}$	196/18	-	87.40	93.63	0.3823	0.1911
		$\mathbf{H_M}$	196/49	-	87.69	93.78	0.3732	0.1866
	MMF	w/o $\mathbf{H}$	196	-	88.66	94.84	0.3097	0.1548
		$\mathbf{H_A}$	196/15	0.3	90.19	95.44	0.2738	0.1369
		$\mathbf{H_T}$	196/18	0.3	90.10	95.44	0.2736	0.1368
		$\mathbf{H_M}$	196/49	0.3	89.92	95.59	0.2645	0.1322
		$\mathbf{H_{A\&T}}$	196/15/18	0.1	90.20	95.72	0.2567	0.1283
		$\mathbf{H_{A\&M}}$	196/15/49	0.1	89.45	95.41	0.2756	0.1378
		$\mathbf{H_{T\&M}}$	196/18/49	0.2/0.1	**90.42**	**95.89**	**0.2468**	**0.1234**
		$\mathbf{H_{A\&T\&M}}$	196/15/18/49	0.05	90.29	95.87	0.2477	0.1239

mance of the subclass classifiers achieves the best performance when the λ_m for different structures is equal, except for $\mathbf{H_{T\&M}}$ in ResNet50. It can be concluded that: 1) For the subclass classifier performance, our MMF model with a single structure is better than the top-down methods with a considerable margin, which verifies the efficiency of the end-to-end training. 2) Besides, MMF with any single structure achieves better performance than "w/o $\mathbf{H}$", indicating the benefit of the superclass classifiers. 3) Furthermore, MMF with multiple label structures performs better than any single one, which confirms our assumption that multiple label structures can provide richer similarity constraints to improve the performance of the subclass classifier. 4) The gain in hierarchical evaluation metrics is more obvious than the flat measure Acc, indicating that predictions in our MMF model are more closer to the ground truth (i.e., a less serious mistake).

5   Conclusion

In this paper, we have constructed a multi-task multi-structure fusion model for hierarchical classification. Various factors have been explored, such as different label structures based on the affinity matrix, the stages to attach the superclass classifiers, and theconstraint intensities. Besides, the hierarchical evaluation metrics have been adjusted to fit the classification with multiple label structures. The experimental results demonstrate that different label structures provide various prior knowledge for the subclass classifier. Meanwhile, integrating these multiple label structures can achieve better results.

In this work, relations of the subclass and its superclasses are impplicitly modeled by the weighted multi-task loss function. In the future, we will explore more direct ways to utilize multiple label structures.

References

1. Ahmed, K., Baig, M.H., Torresani, L.: Network of experts for large-scale image categorization. In: ECCV. pp. 516–532 (2016)
2. Barz, B., Denzler, J.: Hierarchy-based image embeddings for semantic image retrieval. In: WACV. pp. 638–647 (2019)
3. Bengio, S., Weston, J., Grangier, D.: Label embedding trees for large multi-class tasks. In: NIPS. pp. 163–171 (2010)
4. Bertinetto, L., Müller, R., Tertikas, K., Samangooei, S., Lord, N.A.: Making better mistakes: Leveraging class hierarchies with deep networks. CoRR **abs/1912.09393** (2019)
5. Chen, Y., Wang, W., Zhou, Y., Yang, F., Yang, D., Wang, W.: Self-training for domain adaptive scene text detection. In: ICPR. pp. 850–857 (2020)
6. Chen, Y., Zhou, Y., Yang, D., Wang, W.: Constrained relation network for character detection in scene images. In: PRICAI. vol. 11672, pp. 137–149 (2019)
7. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
8. Deng, J., Krause, J., Berg, A.C., Li, F.: Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In: CVPR (2012)
9. Deng, J., Satheesh, S., Berg, A.C., Li, F.: Fast and balanced: Efficient label tree learning for large scale object recognition. In: NIPS. pp. 567–575 (2011)
10. Devries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. CoRR **abs/1708.04552** (2017)
11. Fan, J., Zhou, N., Peng, J., Gao, L.: Hierarchical learning of tree classifiers for large-scale plant species identification. TIP **24**(11), 4172–4184 (2015)
12. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: NIPS. pp. 2121–2129 (2013)
13. Griffin, G., Perona, P.: Learning and using taxonomies for fast visual categorization. In: CVPR (2008)
14. Guillaumin, M., Ferrari, V.: Large-scale knowledge transfer for object localization in imagenet. In: CVPR (2012)
15. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: ICCV Workshops. pp. 554–561 (2013)
16. Lei, H., Mei, K., Zheng, N., Dong, P., Zhou, N., Fan, J.: Learning group-based dictionaries for discriminative image representation. Pattern Recognit. **47**(2), 899–913 (2014)
17. Li, S., Liu, Z., Chan, A.B.: Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In: CVPR Workshops. pp. 488–495 (2014)
18. Lin, D.: Wordnet: An electronic lexical database. CL **25**(2), 292–296 (1999)
19. Liu, B., Sadeghi, F., Tappen, M.F., Shamir, O., Liu, C.: Probabilistic label trees for efficient large scale image classification. In: CVPR (2013)
20. Liu, Y., Dou, Y., Jin, R., Li, R.: Visual confusion label tree for image classification. CoRR **abs/1906.02012** (2019)
21. Liu, Y., Dou, Y., Jin, R., Qiao, P.: Visual tree convolutional neural network in image classification. CoRR **abs/1906.01536** (2019)
22. Luo, D., Fang, B., Zhou, Y., Zhou, Y., Wu, D., Wang, W.: Exploring relations in untrimmed videos for self-supervised learning. CoRR **abs/2008.02711** (2020)

23. Luo, D., Liu, C., Zhou, Y., Yang, D., Ma, C., Ye, Q., Wang, W.: Video cloze procedure for self-supervised spatio-temporal learning. In: AAAI. pp. 11701–11708 (2020)
24. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: NIPS. pp. 849–856 (2001)
25. Qiao, Z., Qin, X., Zhou, Y., Yang, F., Wang, W.: Gaussian constrained attention network for scene text recognition. In: ICPR. pp. 3328–3335 (2020)
26. Qiao, Z., Zhou, Y., Yang, D., Zhou, Y., Wang, W.: SEED: semantics enhanced encoder-decoder framework for scene text recognition. In: CVPR. pp. 13525–13534 (2020)
27. Qin, X., Zhou, Y., Wu, D., Yue, Y., Wang, W.: FC2RN: A fully convolutional corner refinement network for accurate multi-oriented scene text detection. CoRR **abs/2007.05113** (2020)
28. Qin, X., Zhou, Y., Yang, D., Wang, W.: Curved text detection in natural scene images with semi- and weakly-supervised learning. In: ICDAR. pp. 559–564 (2019)
29. Qu, Y., Lin, L., Shen, F., Lu, C., Wu, Y., Xie, Y., Tao, D.: Joint hierarchical category structure learning and large-scale image classification. TIP **26**(9), 4331–4346 (2017)
30. Qu, Y., Wu, S., Liu, H., Xie, Y., Wang, H.: Evaluation of local features and classifiers in BOW model for image classification. MTP **70**(2), 605–624 (2014)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
32. Verma, A., Qassim, H., Feinzimer, D.: Residual squeeze CNDS deep learning CNN model for very large scale places image recognition. In: UEMCON. pp. 463–469 (2017)
33. Wang, Y., Forsyth, D.A.: Large multi-class image categorization with ensembles of label trees. In: ICME. pp. 1–6 (2013)
34. Wang, Y., Hu, Q., Zhu, P., Li, L., Lu, B., Garibaldi, J.M., Li, X.: Deep fuzzy tree for large-scale hierarchical visual classification. ITFS **28**(7), 1395–1406 (2020)
35. Wang, Y., Wang, Z., Hu, Q., Zhou, Y., Su, H.: Hierarchical semantic risk minimization for large-scale classification. ITC (2021)
36. Wu, H., Merler, M., Uceda-Sosa, R., Smith, J.R.: Learning to make better mistakes: Semantics-aware visual food recognition. In: ACM MM. pp. 172–176 (2016)
37. Yang, D., Zhou, Y., Wu, D., Ma, C., Yang, F., Wang, W.: Two-level residual distillation based triple network for incremental object detection. CoRR **abs/2007.13428** (2020)
38. Yao, Y., Liu, C., Luo, D., Zhou, Y., Ye, Q.: Video playback rate perception for self-supervised spatio-temporal representation learning. In: CVPR. pp. 6547–6556 (2020)
39. Zhang, Y., Liu, C., Zhou, Y., Wang, W., Wang, W., Ye, Q.: Progressive cluster purification for unsupervised feature learning. In: ICPR. pp. 8476–8483 (2020)
40. Zhang, Y., Zhou, Y., Wang, W.: Exploring instance relations for unsupervised feature embedding. CoRR **abs/2105.03341** (2021)
41. Zhao, B., Li, F., Xing, E.P.: Large-scale category structure aware image categorization. In: NIPS. pp. 1251–1259 (2011)
42. Zhao, H., Hu, Q., Zhu, P., Wang, Y., Wang, P.: A recursive regularization based feature selection framework for hierarchical classification. ITKDA (2020)
43. Zhao, S., Zou, Q.: Fusing multiple hierarchies for semantic hierarchical classification. IJMLC **6**(1), 47 (2016)