HIERARCHICAL MULTI-LABEL SHIP RECOGNITION IN REMOTE SENSING IMAGES USING LABEL RELATION GRAPHS

Jingzhou Chen, Yuntao Qian

College of Computer Science, Zhejiang University, Hangzhou, China

IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9554687

ABSTRACT

Hierarchical multi-label classification (HMC) aims to assign multiple labels to every instance with the labels organized under hierarchical relations. In the application of ship recognition in remote sensing images, a ship can own coarse-to-fine hierarchical labels, e.g., the military ship, aircraft carrier, and nimitz class aircraft carrier. In this paper, we propose to combine two forms of loss functions to solve the HMC problem based on the neural network. The first probabilistic classification loss is to encode the hierarchical knowledge by introducing hierarchy and exclusion (HEX) graphs to impose constraints on hierarchical labels. The second cross-entropy loss imposes the softmax normalization on leaf nodes in the hierarchy to discriminate fine-grained classes. We evaluate our method on the high resolution satellite image dataset for ship recognition (HRSC), in which hierarchical labels are organized as the three-level tree. The proposed method shows comparative results compared to state-of-art HMC models.

Index Terms— Hierarchical Multi-label Classification, Neural Network

1. INTRODUCTION

Traditional classification methods suppose the labels are mutually exclusive, whereas for hierarchical multi-label classification [1], labels are not disjoined but organized under a hierarchical structure. Such structure can be a tree or a directed acyclic graph (DAG), which indicates the parent-child relations between labels. Every prediction must be coherent, i.e., respect the hierarchy constraint. The hierarchy constraint states that a sample belonging to a given class must also belong to all its ancestors in the hierarchy. HMC problems naturally arise in many domains, such as image classification [2, 3], text categorization [4, 5], and bioinformatics tasks [6, 7].

We study the HMC problem in remote sensing images. Accurate classification of fine-grained ships [8, 9] is vital for numerous civil and military applications. However, fine-grained annotations require expert knowledge and high image quality, and these reasons limit the number of available training samples for each fine-grained class. Although the number of fine-grained images is limited, the prior of the class hierarchy enables us to learn with images only having coarse-grained labels and make coherent predictions.

There are some traditional ways to solve the HMC problem. A straightforward way is to predict labels at the leaf nodes and heuristically add their ancestor labels, which neglects all the non-leaf nodes. Another related way ignores the hierarchy and performs standard multi-label classification, while post-processing is needed to correct label inconsistencies. The hierarchical approaches can be categorized into local and global approaches [1]. Local methods [10] generate a hierarchy of local classifiers used to classify instances following a top-down strategy. Global approaches [11] predict all the levels of classes with a single classifier.

Recent studies usually develop HMC methods based on neural networks. They focus on the design of network architectures [12] or loss functions [13, 14]. [12] employs a cascade of networks, where the layer of each network corresponds to one level of the label hierarchy. Such network architectures generally require all the paths in the label hierarchy to have the same length, which limits their application. Coherent hierarchical multi-label classification networks (C-HMCNN) [13] modifies the standard binary cross-entropy loss by teaching the network when to exploit the prediction of the lower classes in the hierarchy to influence predictions on the upper ones. [14] designs the probabilistic classification loss based on the HEX graph. Considering hierarchical relations, the HEX graph captures three semantic relations between any two labels in the hierarchy: mutual exclusion, overlap, and subsumption. While C-HMCNN only ensures the subsumption relation: the prediction score for a class equals the maximum scores of all its subclasses, including itself.

In this paper, we combine the probabilistic classification loss based on the HEX graph with the cross-entropy loss commonly used in the fine-grained image classification. The main contributions are summarized as follows:

- We exploit the class hierarchy in ship taxonomy to learn a DNN-based classifier with coarse-grained and fine-grained samples. Leveraging the hierarchical label structure, we introduce the HEX graph to design the probabilistic classification loss encoding hierarchical label relations. The multi-class cross-entropy loss is

combined into the loss function to increase the discrimination power of the DNN-based HMC classifier on fine-grained leaf labels.

- We conduct experiments on the real-world ship data set [8], and the experimental results demonstrate that the proposed method is superior to state-of-art HMC methods especially when trained with fewer fine-grained samples.

2. METHOD

In this section, we first describe the network architecture, then explain the corresponding loss function.

2.1. Network architecture

In our experiments, we employ ResNet-50 as the backbone network and replace its output layer with two parallel outputs. Fig. 1 illustrates our network architecture. In Fig. 1, one output contains sigmoid nodes that correspond to all labels in the hierarchy. We adopt sigmoid non-linearity instead of the softmax because sigmoid nodes are independent, whereas the softmax implies mutual exclusion. We constrain their relations by introducing the HEX graph to respect the hierarchy constraint. The HEX graph organizes these sigmoid nodes with three semantic edges: mutual exclusion, overlap, and subsumption. An edge is called an exclusion edge, indicating that two nodes are mutually exclusive, e.g., a ship cannot be the aircraft carrier and merchant ship. If two nodes share no edge, it means that they overlap, i.e., each node can turn on or off without constraining the other. Note that there is no overlap edge in the HEX graph of the experimental ship dataset, as a ship type can not belong to two classes simultaneously. An edge is a hierarchy edge, indicating that one node subsumes another, e.g., the aircraft carrier is a superclass of nimitz class aircraft carrier. Another output includes softmax nodes corresponding to all fine-grained labels in the hierarchy.

A simple demonstration in Fig. 1 explains the class hierarchy existing in the ship dataset, where the aircraft carrier and the merchant ship indicate their respective subclasses, and all classes in the same level of the hierarchy are mutually exclusive. Two classes are mutually exclusive, then all their subclasses are mutually exclusive implicitly. In our network, sigmoid outputs respect such class hierarchy with the aid of the HEX graph. Softmax outputs focus on classifying all fine-grained classes in the class hierarchy. During the inference, softmax scores are added to sigmoid outcomes that correspond to fine-grained labels. Then we obtain combined predictive scores of all labels in the hierarchy.

2.2. Loss Function

The proposed loss function contains two forms of losses. Sigmoid outputs are used to calculate the probabilistic classification loss. Softmax outputs form the cross-entropy loss. Suppose that the number of sigmoid nodes in the HEX graph is n, $\mathbf{y} \in \{0,1\}^n$ is the binary label vector corresponding to all nodes, and $\mathbf{x}$ is the input image, then the joint probability of all nodes is:

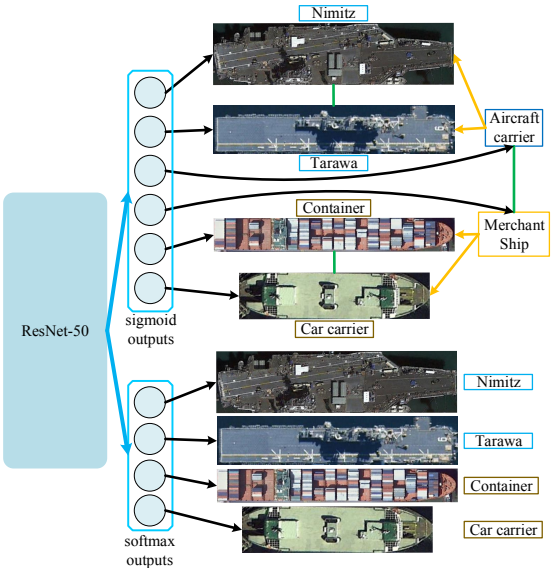

Fig. 1. Our network architecture consists of a backbone network (ResNet-50) and two parallel outputs: sigmoid outputs and softmax outputs. Each node corresponds to a label in the hierarchy with a directed black edge. The directed blue edges indicate features extracted from the backbone network and fed to two outputs. In the HEX graph, directed yellow edges represent subsumption relationships, and undirected green edges stand for mutual exclusion.

$$Pr(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{i}^{n} \phi_i(\bar{x}_i, y_i) \prod \psi_{i,j}(y_i, y_j) \qquad (1)$$

where $\phi_i(\bar{x}_i, y_i) = e^{\bar{x}_i[y_i=1]}$, and $\bar{x}_i$ is the i-th sigmoid output. $\psi_{i,j}(y_i, y_j)$ is the constraint between any two nodes in the HEX graph, defined by:

$$\psi_{i,j}(y_i, y_j) = \begin{cases} 0, & \text{if violates constraints} \\ 1, & \text{otherwise} \end{cases} \qquad (2)$$

$Z(\mathbf{x}) = \sum_{\bar{\mathbf{y}} \in \{0,1\}^n} \prod_i^n \phi_i(\bar{x}_i, \bar{y}_i) \prod_{i,j} \psi_{i,j}(\bar{y}_i, \bar{y}_j)$ sums over all legal binary label vectors and normalizes the joint probability.

Given a training image, it has the observed ground truth label i corresponding to a certain node in the HEX graph, and we can compute the probability of label i by marginalizing all

other labels:

$$Pr(y_i = 1|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \sum_{\bar{\mathbf{y}}:\bar{y}_i=1} \prod_{i}^{n} \phi_i(\bar{x}_i, \bar{y}_i) \prod_{i,j} \psi_{i,j}(\bar{y}_i, \bar{y}_j) \tag{3}$$

In other words, we obtain the marginal probability $Pr(y_i = 1|\mathbf{x})$ of label i by summing over all legal binary label vectors $\bar{\mathbf{y}}$ that include $\bar{y}_i = 1$.

In the training process, labels can be at any level of the hierarchy, and we maximize the marginal likelihood of the observed ground truth labels. Given training examples $\mathcal{D} = \{\mathbf{x}^{(l)}, \mathbf{y}^{(l)}, g^{(l)}\}, l = 1, \ldots, m$, where $\mathbf{y}^{(l)}$ is the ground truth label vector and $g^{(l)} \subseteq \{1, \ldots, n\}$ is the indices of the observed labels, the probabilistic classification loss is:

$$\mathcal{L}_{HEX}(\mathcal{D}) = -\sum_{l} \ln(Pr(y^{(l)}_{g^{(l)}} = 1|\mathbf{x}^{(l)})) \tag{4}$$

To increase the discrimination in leaf nodes (fine-grained classes) of the class hierarchy, we apply the softmax normalization on leaf nodes by adding softmax outputs in the output layer. The softmax normalization increases the probability of the observed ground truth label while decreasing the scores of other labels, which strengthens the discriminative power of the network. We train softmax outputs with the cross-entropy loss and obtain $\mathcal{L}_{CE}$. $\mathcal{L}_{HEX}$ takes advantage of hierarchical knowledge, and $\mathcal{L}_{CE}$ focuses on fine-grained classification. The combined loss is defined as:

$$\mathcal{L}_{com}(\mathbf{x}^{(l)}, y^{(l)}_{g^{(l)}}) = \begin{cases} \mathcal{L}_{CE} + \lambda * \mathcal{L}_{HEX}, & \text{if } g^{(l)} \text{ is in} \\ & \text{leaf nodes} \\ \mathcal{L}_{HEX}, & \text{otherwise} \end{cases} \tag{5}$$

where λ is the hyper-parameter. The total loss on $\mathcal{D}$ is defined as:

$$\mathcal{L}_{total}(\mathcal{D}) = \sum_{l} \mathcal{L}_{com}(\mathbf{x}^{(l)}, y^{(l)}_{g^{(l)}}) \tag{6}$$

3. EXPERIMENTS

We adopt the commonly used dataset HRSC [8] in ship recognition as our experimental dataset. The images sizes range from 300 × 300 to 1500 × 900. The training set contains 617 images, and the test set includes 438 images. There are three levels of hierarchy organized as a tree in HRSC. In the root node, ships and non-ship objects are separated. In the internal nodes, ships are classified into three coarse categories, i.e., aircraft carrier, warcraft, and merchant ship. Leaf nodes in the third level refer to the fine-grained classification where ships are distinguished into their precise categories. In our experiments, we utilize the last two levels of hierarchy that comprise 3 coarse categories and 21 fine-grained subclasses. Note that we crop every ship instance from images according to bounding box annotations to avoid interference from the background.

The whole network is optimized by stochastic gradient descent (SGD) with momentum. Regrading the hyper-parameters, we empirically set the batch size as 32, the momentum as 0.9, and the number of epochs as 40. The initial learning rate is 0.001 and multiplies 0.1 every 20 epochs. The λ is set to 1 according to ablative experiments. The number of samples per class is usually much smaller at the fine-grained level of the hierarchy because of the image quality and expert knowledge. On the other hand, hierarchical knowledge could help learn fine-grained classes. In order to demonstrate the advantage of HMC, we set the following experiments, in which few training samples own fine-grained labels, and some images only have coarse-grained labels. In the training set, we select 0%, 50%, and 90% examples at each fine-grained class and relabel them to their immediate parent classes respectively. All images in the test set are tested with fine-grained labels.

We consider two kinds of evaluation metrics. Note that test images in HRSC contain two levels of hierarchical labels: coarse-grained and fine-grained labels. The first metric only evaluates the fine-grained labels, i.e., fine-grained classification. Given the test image, we assign its fine-grained label with the maximum score in the combined outcomes that correspond to fine-grained labels. The second metric considers both coarse-grained and fine-grained labels. It is computed as the area under the average precision recall curve $AU(\overline{PRC})$, whose points $(\overline{Prec}, \overline{Rec})$ are computed by varying the threshold as: $\overline{Prec} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i}$, $\overline{Rec} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i}$. TP_i, FP_i, and FN_i are the number of true positives, false positives, and false negatives for class i, respectively. $AU(\overline{PRC})$ has the advantage of being independent of the threshold used to predict when a sample belongs to a particular class and is the most used in the HMC literature [6, 12, 13].

We set up two baselines for comparison. The first baseline (softmax-leaf) utilizes ResNet-50, and its softmax output layer corresponds to fine-grained labels. Different from the first one, the softmax output layer in the second baseline (softmax-all) includes all coarse-grained and fine-grained labels. Both baselines are trained with cross-entropy loss. We also compare two state-of-art methods: [14] (HEX) and [13] (C-HMCNN). For a fair comparison, we similarly combine the loss proposed in C-HMCNN with the softmax cross-entropy loss (C-HMCNN-CE). Table 1 records the overall fine-grained classification results (OA) using the first metric and the $AU(\overline{PRC})$ results evaluating HMC methods.

In Table 1, C-HMCNN and our method outperform two baselines when trained with coarse-grained and fine-grained examples. Softmax-all organizes all labels in the hierarchy as mutual exclusive nodes in the softmax output layer. C-HMCNN and our method obtain better OA results, which indicates that the two methods effectively encode hierarchical knowledge. C-HMCNN-CE achieves better OA results

Table 1. OA(%)/$AU(\overline{PRC})$ results on HRSC by relabeling examples at each fine-grained class to their immediate parent classes according to three proportions.

Relabeling	softmax-leaf	softmax-all	HEX	C-HMCNN	C-HMCNN-CE	our
0%	91/	91.7/	88.1/0.96	88.6/**0.98**	91.8/**0.98**	**94.6/0.98**
50%	81.7/	81.5/	51.6/0.74	83.7/0.93	85.8/0.90	**88.2/0.95**
90%	48.8/	42.5/	38.4/0.67	64.3/0.79	72.9/0.72	**75.2/0.90**

than C-HMCNN, which demonstrates that combining with the softmax cross-entropy loss does help improve the discriminative power of the network in fine-grained classification. With the proposed loss function, our method achieves superior $AU(\overline{PRC})$ results compared to the other three HMC methods. Although combined with the softmax cross-entropy loss, C-HMCNN-CE reaches similar $AU(\overline{PRC})$ results to C-HMCNN.

4. CONCLUSION

In this paper, we propose the combined loss function to solve the HMC problem in remote sensing images. With the aid of available hierarchical labels existing in the ship dataset, we evaluate our method for ship recognition. The proposed combined loss function encodes hierarchical knowledge with the probabilistic classification loss by introducing the HEX graph, and it emphasizes fine-grained classification with the softmax cross-entropy loss. Experimental results on HRSC show that the proposed method consistently outperforms baselines and other HMC models under two evaluation metrics.

5. REFERENCES

[1] C. N. Silla Jr and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Mining Knowl. Discovery*, vol. 22, no. 1-2, pp. 31–72, 2011.

[2] I. Dimitrovski, D. Kocev, S. Loskovska, and S. Džeroski, "Hierarchical annotation of medical images," *Pattern Recognit.*, vol. 44, no. 10-11, pp. 2436–2449, 2011.

[3] I. Dimitrovski, D. Kocev, S. Loskovska, and S. Džeroski, "Hierarchical classification of diatom images using ensembles of predictive clustering trees," *Ecological Informat.*, vol. 7, no. 1, pp. 19–29, 2012.

[4] W. Huang *et al.*, "Hierarchical multi-label text classification: An attention-based recurrent network approach," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manag.*, 2019, pp. 1051–1060.

[5] B. Chen, X. Huang, L. Xiao, Z. Cai, and L. Jing, "Hyperbolic interaction model for hierarchical multi-label classification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 7496–7503.

[6] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel, "Decision trees for hierarchical multi-label classification," *Mach. Learn.*, vol. 73, no. 2, pp. 185, 2008.

[7] R. Cerri, R. C. Barros, A. C. P. L. F. de Carvalho, and Y. Jin, "Reduction strategies for hierarchical multi-label classification in protein function prediction," *BMC Bioinformat.*, vol. 17, no. 1, pp. 373, 2016.

[8] Z. Liu *et al.*, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. 6th Int. Conf. Pattern Recognit. Appl. Methods*, 2017, pp. 324–331.

[9] X. Zhang, Y. Lv, L. Yao, W. Xiong, and C. Fu, "A new benchmark and an attribute-guided multilevel feature representation network for fine-grained ship classification in optical remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1271–1285, 2020.

[10] Z. Zou, S. Tian, X. Gao, and Y. Li, "mldeepre: Multifunctional enzyme function prediction with hierarchical multi-label deep learning," *Frontiers Genet.*, vol. 9, pp. 714, 2019.

[11] S. Gopal and Y. Yang, "Recursive regularization for large-scale classification with hierarchical and graphical dependencies," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 257–265.

[12] J. Wehrmann, R. Cerri, and R.C. Barros, "Hierarchical multi-label classification networks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5075–5084.

[13] E. Giunchiglia and T. Lukasiewicz, "Coherent hierarchical multi-label classification networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33.

[14] J. Deng *et al.*, "Large-scale object classification using label relation graphs," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 48–64.