PROGRESSIVE TRAINING ENABLED FINE-GRAINED RECOGNITION

Bin Kang *Fan Wu* *Xin Li* *Quan Zhou*

School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, China.

2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897221

ABSTRACT

Organizing training samples in a meaningful order is beneficial for accelerating the convergence rate and enhancing the recognition performance in the CNN model. However, achieving reasonable sample ranking for fine-grained recognition datasets is very challenging because the intra and inter class relation in those datasets is opposite to that in public recognition datasets. In this paper, we propose a general framework for the progressive training of fine-grained recognition models. In particular, we first formulate the training subset selection as a group ranking-oriented submodular optimization problem, where the submodularity is adopted to evaluate the benefit of selected training subsets. This can give theoretical guidance for the consecutive discrimination of difficult and ordinary training subsets. Secondly, we design a training strategy to dynamically adjust the ratio of difficult and ordinary training subsets according to the recognition performance. Extensive experiments on CUB-200-2011 and Stanford Dogs datasets demonstrate that the proposed method outperforms the state-of-the-art curriculum learning methods.

Index Terms— Fine-grained recognition, Submodular optimization, Group ranking.

1. INTRODUCTION

Fine-grained image recognition, aiming at distinguishing sub-categories of the same basic-level category, involves more challenging factors than traditional classification tasks due to the fact that there only exists subtle difference within sub-categories. To overcome challenges in fine-grained classification, existing works[1–4] focus on the design of the network structure to simultaneously locate and represent the discriminative image sub-regions. In fact, besides network designing, studying the training strategies to organize the training samples in a meaningful order is also very helpful for enhancing the performance gain[5]. However, there has few fine-grained recognition works on this topic.

Curriculum learning(CL) is the representative method that can organize the training samples from easy to hard. Designing the difficulty score of training samples is the key point in curriculum learning. According to the design of difficulty score, the state-of-the-art curriculum learning methods can be roughly categorized into pre-defined methods[6–8] and automatic methods[9–16]. Pre-defined methods manually organize the samples from easy-to-hard using human annotators, i.e. the human response time has been widely used as difficulty scores because it is closely related to a gradation of the visual search task. In contrast, automatic methods automatically estimate difficulty scores by the feedback from the CNN output. This design is derived from an observation that the difficulty of a sample is proportional to the value of CNN loss function.

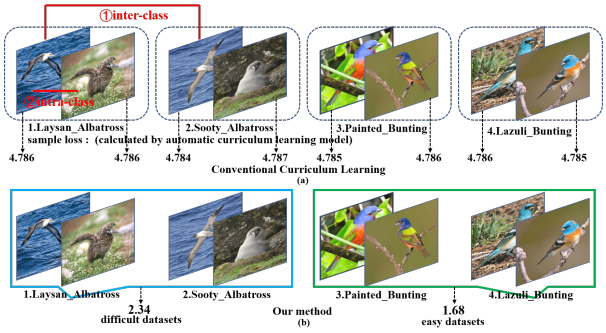

Fig. 1. Illustrating the difference between conventional curriculum learning and our method in fine-grained dataset. Fine-grained dataset is challenging because the intra-class relation is large and the inter-class relation is small. The difficulty scores in conventional curriculum learning lose its efficiency in this dataset.

Although curriculum learning own the potential advantage in improving the generalization of recognition models[5], it cannot be applied directly in fine-grained recognition. The key reason has been clearly shown in Fig. 1(a). Specifically, it does not just contain mature bird samples, the nestling may be also involved in the same class. This will cause a large intra-class variation. In comparison, the inter-class variation between the challenging sub-classes is small. Based on this observation, in the fine-grained classification dataset, there exists a complex training sample relationship, i.e. the intra and inter class relation in challenging sub-classes is opposite to that in ordinary sub-classes. The difficulty scores in

This work was partly supported by the NSFC (Nos. 62171232 61876093). China Postdoctoral Science Foundation under Grant 2020M681684, Jiangsu Postdoctoral Science Foundation under Grant 2021K278B. The Key Research and Developement Plan of Jiangsu Province (Grant No. BE2020729)

traditional curriculum learning methods could not give informative advice because they consider each training sample as an independent individual. The ranking results of traditional curriculum learning will be inevitably confused by high intra-class relation.

In this paper, we do not consider each training sample as an independent individual. Instead, we plan to explore the inter and intra class relation to divide the challenging and ordinary sub-classes into different groups for designing group-oriented training strategy. In particular, we first formulate the difficult group selection as a submodular optimization, where the difficult groups that owe high intra-class relation and serious occlusion can be discriminated through the optimization of the group difficulty indicator. Fig. 1(b) is an example to show the final group scores after submodular optimization, which can give high score to difficult groups while low score to ordinary groups. Based on submodular optimization, we secondly design the training strategy to effectively utilize group ranking results to achieve progressive training.

The main contributions are listed as follows:

(1) We are the first to formulate the challenging sub-class selection as a submodular optimization problem, in which the proposed submodularity-oriented objective function can theoretically guide the consecutive discrimination between difficult and ordinary training subsets.

(2) We integrate the submodular optimization into a progressive training strategy. With the support of submodular optimization, the proposed training strategy can dynamically adjust the ratio of difficult and ordinary training subsets according to the learning state of the recognition model.

(3) Extensive experiments have shown that the proposed submodular optimization based training strategy can be extended to different kinds of fine-grained recognition models.

2. PROPOSED METHOD

2.1. A brief introduction to submodular optimization

Given a finite set $V = \{1, 2, ..., n\}$, a set selecting function f is submodular if it meets

$$f(A \cup \{u\}) - f(A) \geq f(B \cup \{u\}) - f(B) \tag{1}$$

where $A \subseteq B \subseteq V$ and $u \notin B$. We can better understand Eq. (1) through Fig. 2. In Fig. 2(a), if we add D' to set $\{D_1\}$, the increased coverage area can be represented as A. In comparison, in Fig. 2(b), if we add subset D' to set $\{D_1, D_2\}$, the increased coverage area can be represented as B. It can be clearly seen that A is bigger than B. Thus, the submodularity means that: with the increase of subset, the benefit (i.e. increased coverage area) of the new added subset is linearly decreased. Based on submodularity, the submodular optimization can utilize a linear optimization strategy to evaluate the benefit of selecting a series of subsets. In general, the submodular optimization problem with cardinality parameter k is formulated as:

$$\max_{D \subseteq V, |D| = k} f(D). \tag{2}$$

where, D is the selected subset.

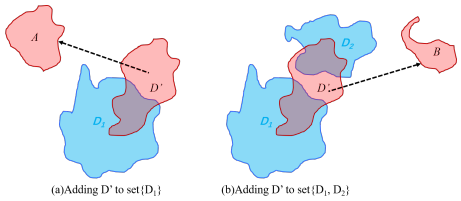

Fig. 2. Illustration the concept of submodularity

2.2. Submodular optimization based difficult group selection

In our training strategy, the key point is to progressively divide the entire datasets into groups from easy to hard. Here we plan to formulate an optimization problem with the goal of picking out difficult groups. Selecting difficult groups in a fine-grained recognition dataset is a tough task because it is required to define a difficulty indicator to explore the relationship between training subsets. It has been investigated in the traditional curriculum learning methods that challenging factors such as high target pose variation and serious occlusion/background clutter are the main characteristic of difficult samples. Beside of above challenges, samples in fine-grained recognition datasets have additional challenging factors: high intra-class variation and low inter-class variation. Thus it is very hard to use a difficulty indicator to evaluate a single sample directly. In this paper, we return to the original but efficient idea that exploring the sample relation to design group difficulty indicator.

2.2.1. Group difficulty indicator

Inspired by fisher criterion, the group difficulty indicator (GDI) is defined as follow

$$GDI = \frac{Ave\_inter}{Ave\_intra} \tag{3}$$

The numerator "*Ave_inter*" is denoted as the average inter-class similarity between different categories, the denominator "*Ave_intra*" is denoted as the average value of intra-class similarity of a certain category. Choosing "*Ave_inter*" as the numerator means that in the optimization process, if some sub-categories are very similar, "*Ave_inter*" can be optimized to obtain a higher score. Similarly, the denominator means that if some sub-categories have high intra-class similarity, "*Ave_intra*" can be optimized to give a lower score for GDI. The index GDI in Eq. (3) is proportional to the difficulty degree of a certain category, which meets the characteristic of fine-grained recognition task.

2.2.2. *Submodular optimization*

Here we define a group is composed of at least one category. Based on this definition, the process of selecting difficult categories is formulated as a *GDI* indicator based optimization. The submodular optimization function is expressed as follows:

$$\max_{|D|=m} \frac{1}{|D| \cdot |S|} \sum_{\mathbf{X}_j \in S} f(\mathbf{X}_i, \mathbf{X}_j) - \lambda \frac{1}{|D|} \sum_{\mathbf{X}_i \in D} \frac{2}{|\mathbf{X}_i| \cdot (|\mathbf{X}_i| - 1)} \sum_{u,v \in X_i} f(\mathbf{x}_i^u, \mathbf{x}_i^v) \tag{4}$$

where D denotes the set of categories are to be selected, S denotes the set of categories that are similar to D, f is the selecting function, which is submodular. $\mathbf{X}_i$ denotes the ith category in D, $\mathbf{X}_j$ denotes the jth category in S. $\mathbf{X}_i = [\mathbf{x}_i^1, ..., \mathbf{x}_i^u, ..., \mathbf{x}_i^n]$, n denotes the number of samples in $\mathbf{X}_i$. $|D| = m$ is the cardinality constraint for selecting m categories.

The optimization problem Eq. (4) is aimed to select a subset D with m samples from N categories. The advantage of this design is that we can make the selected subsets D become more discriminative through maximizing the inter-class similarity and minimizing the intra-class similarity. Specifically, the first term $\sum_{\mathbf{X}_j \in S} f(\mathbf{X}_i, \mathbf{X}_j)$ in Eq. (4) indicates the average inter-class similarity between similar categories. The second term of Eq. (4) indicates the average value of intra-class similarity of difficult samples in a certain category. Parameter $\lambda \geq 0$ balances the importance between intra-class and inter-class similarity.

Eq. (4) is a non-monotonic submodular optimization with cardinality constraints. Inspired by the work in [17], we propose a random greedy algorithm, which can reach an approximation of 1/e for general non-monotone objectives. The time complexity of proposed algorithm is O(nm). Details of the algorithm are shown in Algorithm 1.

Algorithm1 Random Greedy Algorithm

Input: $D_0 \leftarrow \emptyset$: initialize the difficulty category, N:total category
Output:D_m:a set contain m difficult categories
1: **for** $i = 1$ to m **do**
2: Let $M_i \subseteq N \setminus D_{i-1}$ be a subset of size m maximizing $\sum_{u \in M_i} h(u \cup D_{i-1}) - h(D_{i-1})$
3: Let u_i be a uniformly random element from M_i
4: Let $D_i \leftarrow D_i + u$
5: **end for**

2.3. Progressive training strategy

Traditional CL is the sample based training strategy. In contrast, our design is a group based training strategy. Thus traditional CL methods could not give much information. In our training strategy, the main innovation is to define a simple yet efficient parameter namely α to indicate the ratio of difficult and ordinary subsets. As the value of α dynamically decreases, the classifier focuses on ordinary training subsets at first, gradually focuses on difficult training subsets later. Based on this design, we successfully integrate the difficult group selection into a progressive training framework (see Fig. 3).

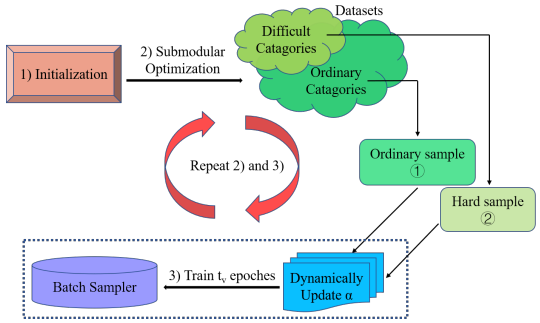

Fig. 3. The flow chart of our progressive training strategy.

3. EXPERIMENTS

In this section, we first briefly introduce the experimental setting including datasets and implement details. Then we test the generality of our proposed progressive training strategy, where representative CNN backbones and state-of-the-art fine-grained recognition methods are included. Finally, we carry out an ablation study to test the effectiveness of submodular optimization based difficult group selection.

3.1. Experimental setting

Datasets: We explore the effectiveness of our proposed progressive training strategy on two fine-grained datasets, i.e., CUB-200-2011[18] and Stanford Dogs[19].
Implement details: For fair comparison, an input image is firstly resized to 600 * 600 and then it is cropped into 448 * 448 for CUB-200-2011 and Stanford Dogs(random cropping for training and center cropping for testing). The batch size is set to 64. We select the SGD optimizer to optimize the classifier with a momentum of 0.9. The learning rate is initialized as 0.01 for CUB-200-2011 and 0.003 for Stanford Dogs. We implement the experiments on four Nvidia Tesla P100 GPUs.

3.2. Experimental results

Experiment on public datasets: In this test, we show the experimental results on CUB-200-2011 and Stanford Dogs datasets (see Table. 1). The original recognition model that does not contain training strategy is named as **w/o strategy**. The model that involves our progressive training strategy is named as **w strategy**. From Table 1 we could clearly see that after adding our progressive training strategy to three

Table 1. Comparison of different methods on two fine-grained recognition datasets.

Method	CUB-200-2011		Stanford Dogs	
	w/o strategy(%)	**w** strategy (%)	**w/o** strategy(%)	**w** strategy(%)
VGG16[20]	73.4	**74.25**	68.32	**71.47**
ResNet50[21]	82.39	**83.41**	84.69	**85.69**
DesNet121[22]	80.79	**81.81**	79.9	**80.96**
API-Net[23]	87.45	**88.16**	79.9	**80.5**
CAL[24]	89.98	**90.47**	88.96	**89.45**

CNN baselines, they achieves about a 1.1% improvement in CUB-200-2011. In Stanford Dogs datasets, our training strategy again gets 1.2% and 0.6% performance gain in CNN backbone and a top fine-grained recognition method namely API-Net. This greatly shows the generality of our proposed training strategy.

Comparison to curriculum learning: The competitors in this test are transfer teacher[11] and self-paced learning[9], which are two well-known curriculum learning methods. The testing results are shown in Fig. 4. From Fig. 4 we could clearly see that adding the proposed training strategy to ResNet50 can give faster convergence and higher accuracy than two curriculum learning methods.

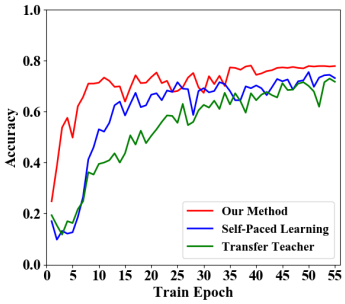

Fig. 4. Experimental comparison to curriculum learning. Our method (in red) can reach higher accuracy and faster convergence than curriculum learning methods in fine-grained datasets.

Visualization: To validate the effectiveness of our method, the Grad-CAM method[25] is used to generate the heatmaps for visualizing the performance of progressive training strategy. The visual results are shown in Fig. 5. From Fig. 5 we could see that the integration of difficult group selection and the progressive training can guide the CNN backbone to locate informative sub-regions of difficult sample.

3.3. Ablation study

Submodular optimization is the core in our progressive training strategy. In this part, we will show the efficiency of proposed submodular optimization. Specifically, it has been verified in curriculum learning that the difficulty of image is proportional to the value of CNN loss function. This means

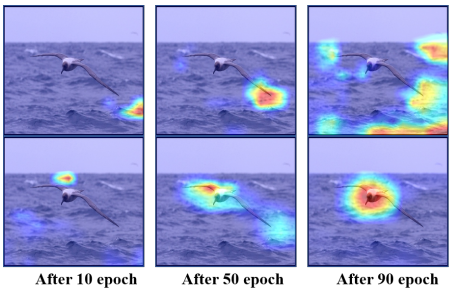

Fig. 5. Visualization results. The first row show the result on ResNet50 backbone without any training strategy while the second row show ResNet50 backbone with our training strategy.

that the accuracy of testing model can be considered as an indicator to test the ranking results. Based on this observation, we use the proposed submodular optimization to rank the groups, the pre-trained ResNet50 is used to calculate the averaged accuracy of different groups. From Fig. 6 we could clearly see that the averaged group recognition accuracy follows an increasing trend, which can verify the efficiency of group ranking result.

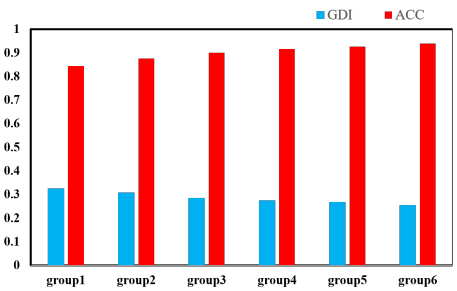

Fig. 6. Testing the efficiency of the proposed submodular optimization. The proposed submodular optimization is derived from group difficulty indicator (*GDI*). As *GDI* decreases, the averaged accuracy of the selected group gradually increases, indicating that the proposed submodular optimization can give a reasonable ranking result.

4. CONCLUSION

In this letter, we propose to integrate the difficult group selection and the progressive training into a submodular optimization framework, which progressively generates training subsets to guide the training process in a meaningful order. Extensive experiments have shown that our method can be extended to various fine-grained recognition models with prominent accuracy enhancement.

References

[1] R.Du, D.Chang, and A.K.Bhunia et.al, "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *ECCV*, 2020.

[2] W.Ge, X.Lin, and Y.Yu, "Weakly supervised complementary parts models for fine-grained image classification from the bottom up," in *CVPR*, 2019.

[3] T.Y.Lin, A.RoyChowdhury, and S.Maji, "Bilinear cnn models for fine-grained visual recognition," in *ICCV*, 2015.

[4] J.He, J.N.Chen, and S.Liu et.al, "Transfg: A transformer architecture for fine-grained recognition," in *CVPR*, 2021.

[5] Y.Chen X.Wang and W.Zhu, "A survey on curriculum learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence(Early Access)*, March 2021.

[6] R.Tudor Ionescu, B.Alexe, and M.Leordeanu et.al, "How hard can it be? estimating the difficulty of visual search in an image," in *CVPR*, 2016.

[7] P.Soviany, C.Ardei, and R.T.Ionescu et.al, "Image difficulty curriculum for generative adversarial networks," in *WACV*, 2020.

[8] R.El-Bouri, D.Eyre, and P.Watkinson et.al, "Student-teacher curriculum learning via reinforcement learning: Predicting hospital inpatient admission location," in *PMLR*, 2020.

[9] K.Ghasedi, X.Wang, and C.Deng et.al, "Balanced self-paced learning for generative adversarial clustering network," in *CVPR*, 2019.

[10] G.Hacohen and D.Weinshall, "On the power of curriculum learning in training deep networks," in *PMLR*, 2019.

[11] D.Weinshall, G.Cohen, and D.Amir, "Curriculum learning by transfer learning: Theory and experiments with deep networks," in *PMLR*, 2018.

[12] A.Oliver T.Matiisen and T.Cohen et al, "Teacher-student curriculum learning," *IEEE Transactions on Neural Networks and Learning Systems*, September 2019.

[13] Y.P.Tang and S.J.Huang, "Self-paced active learning: Query the right thing at the right time," in *AAAI*, 2019.

[14] Y.Wang, W.Gan, and J.Yang et.al, "Dynamic curriculum learning for imbalanced data classification," in *ICCV*, 2019.

[15] J.Han D.Zhang and L.Zhao et al, "Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework," *International Journal of Computer Vision*, August 2019.

[16] M.Zhao, H.Wu, and D.Niu et.al, "Reinforced curriculum learning on pre-trained neural machine translation models," in *AAAI*, 2020.

[17] N.Buchbinder, M.Feldman, and J.Naor et.al, "Submodular maximization with cardinality constraints," in *SODA*, 2014.

[18] C.Wah, S.Branson, and P.Welinder et.al, "The caltech-ucsd birds-200-2011 dataset," in *CVPR*, 2011.

[19] A.Khosla, N.Jayadevaprakash, and B.Yao et.al, "Novel dataset for fine-grained image categorization: Stanford dogs," in *CVPR*, 2011.

[20] K.Simonyan and A.Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] K.He, X.Zhang, and S.Ren et.al, "Deep residual learning for image recognition," in *CVPR*, 2016.

[22] G.Huang, Z.Liu, and L.Van Der Maaten et.al, "Densely connected convolutional networks," in *CVPR*, 2017.

[23] P.Zhuang, Y.Wang, and Y.Qiao, "Learning attentive pairwise interaction for fine-grained classification," in *AAAI*, 2020.

[24] Y.Rao, G.Chen, and J.Lu et.al, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *ICCV*, 2021.

[25] R.R.Selvaraju, M.Cogswell, and A.Das et.al, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.