Leveraging Fine-Grained Labels to Regularize Fine-Grained Visual Classification

Junfeng Wu
College of System Engineering
National University of Defense Technology
Changsha 410073, China
+86 18274899915
wjf_jlro@126.com

Li Yao
College of System Engineering
National University of Defense Technology
Changsha 410073, China
+86 13755130950
liyao6522@gmail.com

Bin Liu
College of System Engineering
National University of Defense Technology
Changsha 410073, China
+86 15116333743
liubin11@nudt.edu.cn

Zheyuan Ding
College of System Engineering
National University of Defense Technology
Changsha 410073, China
+86 15116333743
371865833@qq.com

ABSTRACT

Fine-grained visual categorization (FGVC) is challenging mainly due to the large intra-class confusion and small inter-class variance in terms of shape, pose, and appearance. We propose the concept of fine-grained label and that any given label can be further classified into some sub-classes as fine-grained labels, and thus samples of each original label are classifed into several sub-classes in which only more familiar samples are given the same fine-grained label. The samples of fine-grained labels have less intra-class confusion and bigger inter-class variance. Besides, fine-grained labels can be obtained through unsupervised means without any domain knowledge or annotations. Instead of training on the fine-grained labels directly, we utilize these "free" labels as an auxiliary task to regularize the training of the deep learning model. In the test phase, as sub-classes of the original label, the predicted fine-grained labels are used for integration with original labels to get the final classification results. Experiments on the popular CUB-200-2011 dataset demonstrate that employing the proposed fine-grained labels in CNN model improves performance from both training and test phases.

CCS Concepts

Computing methodologies → Machine learning → Learning paradigms→Multi-task learning.

ICCMS 2019, January 16–19, 2019, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-6619-9/19/01…$15.00
DOI: https://doi.org/10.1145/3307363.3307382

Keywords

fine-grained visual categorization; fine-grained label; regularize; integration

1. INTRODUCTION

Fine-grained visual categorization, which aims to distinguish among subordinate categories with images or videos, such as identifying car models or discriminating bird species, has received increasing attention in recent years [4, 7, 9, 17]. Compared to generic visual categorization, this task is more challenging since the subtle visual differences can be easily overwhelmed by other factors such as poses, illumination or viewpoints.

Humans typically distinguish subordinate categories according to the difference in some semantic parts, and differences between classes can only be found at some local area or parts in FGVC. Most modern methods [12, 13, 16] for FGVC rely on a combination of localizing discriminative regions or parts and learning corresponding discriminative features. However, these methods usually require strong supervision such as keypoint or attribute annotations, which are quiet difficult and expensive to be obtained at scale.

Taxonomy or given class hierarchies have been applied for image classification as well as FGVC [1, 2, 11]. Such methods use taxonomy or given class hierarchies to generate auxiliary tasks to regularize the training of primary task, and they do not need annotations for each sample. However, these methods require domain knowledge to utilize the relations between labels, such as parent-child and sibling-sibling relations.

In this work, we propose the concept of fine-grained label, which is the sub-class of original label. Each original label can be classified into several fine-grained labels, so that only more similar samples will be labeled with the same fine-grained label. The samples of each fine-grained label are even less than that of the original labels, but the fine-grained labels are used to generate an auxiliary task to regularize the training of the deep learning model. Besides, as sub-class of the original label, the predicted fine-grained label can be used for integration with the original label to get a more robust result. The fine-grained labels are

obtained through unsupervised clustering and thus require no prior or annotation. We note that our method can be used in conjunction with more complex models such as PFNet[11] to further improve performance for FGVC.

The rest of the paper is organized as follows: Section 2 present the most closely related work. Section 3 is devoted to the overview of our method and how the proposed fine-grained labels are used for FGVC, while the experimental details are articulated in Section 4. Conclusions and future directions of our work are given in Section 5.

2. RELATED WORK

The idea of fine-grained label comes from the taxonomy based image classification, which uses the parent or sibling class of the original subordinate class of FGVC. Taxonomy based image classification is usually related to multi-task learning which merely use auxiliary task to regularize the training of deep learning model, however, in our work the fine-grained label can be used for both training and test.

Unlike ordinary image classification tasks, the differences between classes are more subtle in FGVC, in which images have very small signal-to-noise ratio and some categories can only be distinguished by small local differences. Therefore, how to find and make full use of these useful local information becomes the key to the success of FGVC algorithm. At present, most classification algorithms [5-7] follow such a framework: first find the foreground object and its local area which include the crucial parts of the object, and then extract the features of these areas. After proper processing of the obtained features, they are used to complete the training and prediction of the classifier. Some studies have shown that the features extracted from the deep convolution neural network have more powerful description ability than the artificial features. The application of the deep convolution features to the FGVC task can achieve better results [3]. The addition of depth convolution features brings new opportunities for the development of FGVC.

In recent years, more and more studies have tended not to use these annotation information, relying only on category labels to complete image classification tasks [1, 8, 15]. Taxonomy based image classification is one of this direction and the most related work is [1], which uses label hierarchy to generate auxiliary tasks to regularize the deep learning algorithm. However, it require taxomomy or domain knowledge as prior and the auxiliary tasks is only used to get a more sophisticated loss function in the training phase. Our proposed fine-grained labels are derived from original label through unsupervised clustering in training and the predicted fine-grained labels of test samples can be used for integration to get a more accurate prediction.

3. THE PROPOSED METHOD

3.1 Overview

The tasks or the labels in the traditional multi-task learning are usually related, while in our framework the auxiliary task based on fine-grained labels are more than just related to the primary task. Since each fine-grained label derive from its own original label, the derived fine-grained labels can help determine the original label in the test phase, and thus this auxiliary task can not only help to train the parameters of the model to gain generalized features but also can be used for integration to get the final classification results. Figure 1 shows the architecture of our proposed Fine-grained Label Assistanting deep learning model.

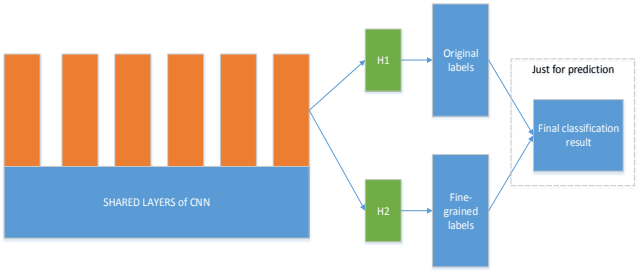

Figure 1. The architecture of our proposed Fine-grained Label Assistanting deep learning model. H_i means the hidden layer and it is to be noted that orginal labels are integrated with fine-grained labels in the test phase for the final prediction.

We aim to optimize the performance of a main or primary task T_p with the aid of the additional auxiliary task T_a and the general form of the objective function that we aim to minimize here is:

$$\arg\min_{Wp,Wa} \sum_{i=1}^{N} [l_p(y_i^p, f_p(x_i, W_p)) + \alpha l_a(y_i^a, f_a(x_i, W_a))] \quad (1)$$

The index "p" and "a" refers to the primary task and the auxiliary task respectively. For N input samples, W_t denotes the weights of the network with respect to task t ($t=p$ or a), while y_t^i denotes the ground truth for the input representation x_i . f_t represents the feature transformation of the input x_i with respect to the task t and the corresponding weights W_t , and l_t is the corresponding loss function for the task. It is to be noted that $W_t = (W_s, W_{tt})$, where W_s is the shared representation, *i.e.*, the weights of the shared layers, while W_{tt} is the set of weights of the task specific layers. We associate the auxiliary task with its loss function l_t and a coefficient α, which acts as a weight determining the relative importance of the auxiliary task in training.

3.2 Fine-grained Label

Any given label can be classified in a more fine-grained fashion, and we will naturally think of some semantic-related features such as breed, age, sex and so on. Although these semantic-related features can be useful, they are more unavailable than the additional expensive annotations such as bounding box or keypoints. We propose a clustering based method to obtain the fine-grained labels. The steps are:

1）Training the CNN model with original labels
2）Obtaining the softmax value for each sample of each original label
3.）Clustering with these softmax values for each original label seperately and obtain fine-grained labels

The fine-grained labels obtained in this way are actually some cluster numbers and it might be difficult to give these clusters some meaningful names, but it will not influence how we use them in the training or test phase. In the test phase, the predicted fine-grained labels are used for integration of final prediction as follows:

$$y_{pred} = \arg\max_i [s(i) + \beta s(SF(i))] \quad (2)$$

[1] The code of PFNet is available at https://github.com/MichaelLiang12/PFNet-FGVC.

Where y_{pred} is the final prediction for a test sample, and i refers to the original label. $s(i)$ is the softmax value of the test sample for label i, and $SF(i)$ is the set of the fine-grained labels for original label i. β is the coefficient which acts as a weight determining the relative importance of the auxiliary task for prediction. In this way, the final prediction take both kind of labels into consideration, which makes the final result more robust.

Although dividing the original labels of FGVC problem into fine-grained ones makes samples of each new label even less, the loss function is a combination which takes both original and fine-grained labels into consideration. These fine-grained labels are used to learn the subtle difference within the original label, and thus obtaining more fine-grained features.

From Figure 2, it is evidently shown what are large intra-class confusion and small inter-class variance, and the three rows of the image represent three kinds of fine-grained labels of California Gull and Western Gull respectvely. We can see that each sample of new fine-grained labels is partially representive of its own original label. If we do not differentiate them with fine-grained labels, these confusing samples can hardly be telled from each other since subtle visual differences between similar classes can be easily overwhelmed by differences within one class.

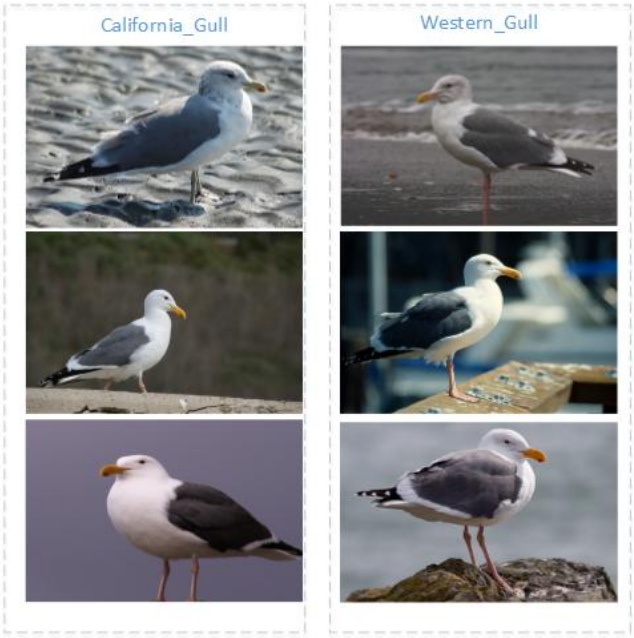

Figure 2. Samples belonging to different fine-grained labels of California Gull and Western Gull in Caltech-UCSD Birds-200-2011.

4. EXPERIMENTS

Experiments are carried out on the widely used FGVC dataset Caltech-UCSD Birds-200-2011. The main purpose of the experiments is to verify whether the proposed fine-grained labels can be used to improve the classification performance. PyTorch is used to conduct all experiments on an NVIDIA GTX 1080 Ti GPU.

4.1 Dataset

The Caltech-UCSD Birds-200-2011 dataset contains 11788 images of birds belonging to 200 classes, and the training and test splits are roughly equal in size (5994 vs. 5794). Since most species have 30 samples for training, we divide each original label into a fixed number of n fine-grained labels for simplicity.

4.2 Experimental Setup

We use the Imagenet pre-trained model vgg19 to initialise our network. We take all the layers till the last 4096-dimensional fully connected layer. Right after this layer, we create 2 branches. Both branches have a linear layer mapping the 4096 dimensional feature of the shared layer to 4096 dimensions, followed by ReLU activation function and dropout with a probability of 0.5. Finally, a linear layer connects the 4096 dimensional feature to the corresponding number of classes. As a common way of data augmentation, we choose ramdomly 224*224 RGB image from original images and flip them horizontally as input. Mini-batch gradient descent is used for training, with a batch size of 32 and an initial learning rate of 0.001 which decays every 10 epochs.

The nunber of fine-grained labels for each original label is set to 3. Since softmax operation always get a vector in which only one element is big enough to 1, K-medoids [10] which diminish the sensitivity to outliers is used for clustering with the softmax value of each sample in the training phase to get the fine-grained labels.

To find the optimal value for α and β, we first set $\beta = 0$ when the framework degrades into an ordinary multi-task learning, and we find that α=0.3 is the optimal choice. After setting α, we find the globally optimal value 0.8 for β.

4.3 Experimental Results

We refer the original vgg19 model as vgg19 Single Task model (vgg19-ST), while the model with fine-grained labels as an auxiliary task as Fine-grained Label Assistanting model (vgg19-FLA). If β is set to 0 in FLA, the model degenerates into an oridinary multi-task learning with fine-grained labels. From table I, it is evident that the performance of vgg19 is improved with fine-grained labels as an auxiliary task to gain more generalized features in the training phase, and the performance is further improved with predicted fine-grained labels intergrated with the predicted original label for final prediction.

We further implement the PFNet model of [8], And refer it to PFNet-FLA model. It is shown that combining a fine-grained label based classifier with the PFNet model gives a slight boost in accuracy. While performance of our Fine-grained Label Assistanting model does not reach state of the art, it surpasses [1, 5, 6, 14], which rely on expensive additional annotations of keypoints or taxonomy.

Table 1. The comparision of accuracy between the original FGVC algorithms and the ones combined with FLA.

Approach	Accurary
vgg19-ST	76.87
vgg19-FLA (β = 0)	77.41
vgg19-FLA	77.95
PFNet	84.32
PFNet-FLA	84.86

[10]	76.66
[15]	75.04
[16]	75.73
[17]	73.89

From Figure 3, it can be seen that our proposed vgg19-FLA model get a more accurate result than vgg19-ST model, while the rate of convergence is about the same to vgg19-ST model.

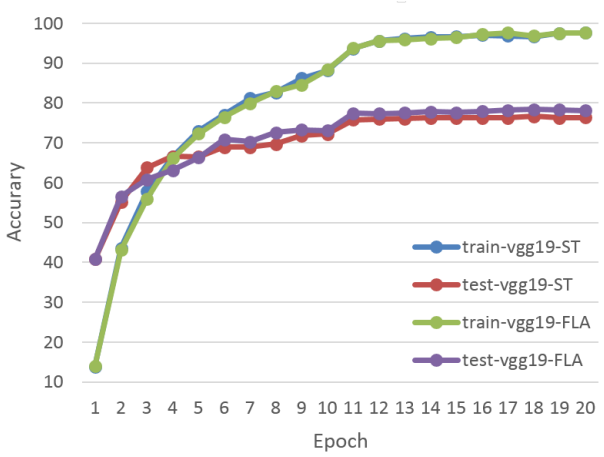

Figure 3. The comparision of vgg19-FLA with vgg19-ST in the training and test phase.

5. CONCLUSION

We propose the concept of fine-grained label and the way to improve FGVC with it in both training and test phase. Besides, we need no prior or semantic annotation to obtain fine-grained label. Experiments on the dataset Caltech-UCSD Birds-200-2011 demonstrate the effectiveness of our proposed FLA model.

In the near future, we will test our method with more datasets and learn the coefficient α and β for each class through reinforce learning instead of fixing them to set values.

6. ACKNOWLEDGMENTS

Our thanks to ACM SIGCHI for allowing us to modify templates they had developed. The presented work is framed within the National Natural Science Foundation of China (No. 71371184). We also thank Bin Liu and Zheyuan Ding for giving us revision advice.

7. REFERENCES

[1] DASGUPTA, R. and NAMBOODIRI, A.M., 2017. Leveraging multiple tasks to regularize fine-grained classification. In *International Conference on Pattern Recognition*, 3476-3481.

[2] DENG, J., DING, N., JIA, Y., FROME, A., MURPHY, K., BENGIO, S., LI, Y., NEVEN, H., and ADAM, H., 2014. Large-scale object classification using label relation graphs. In *European conference on computer vision* Springer, 48-64.

[3] DONAHUE, J., JIA, Y., VINYALS, O., HOFFMAN, J., ZHANG, N., TZENG, E., and DARRELL, T., 2013. A deep convolutional activation feature for generic visual recognition. arXiv preprint. *arXiv preprint arXiv:1310.1531*.

[4] FU, J., ZHENG, H., and MEI, T., 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Cvpr*, 3.

[5] GAO, Y., BEIJBOM, O., ZHANG, N., and DARRELL, T., 2016. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 317-326.

[6] HE, K., ZHANG, X., REN, S., and SUN, J., 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European conference on computer vision* Springer, 346-361.

[7] HUANG, S., XU, Z., TAO, D., and ZHANG, Y., 2016. Part-stacked cnn for fine-grained visual categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1173-1182.

[8] KONG, S. and FOWLKES, C., 2017. Low-rank bilinear pooling for fine-grained classification. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on* IEEE, 7025-7034.

[9] LIN, T.-Y., ROYCHOWDHURY, A., and MAJI, S., 2015. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 1449-1457.

[10] PARK, H.-S. and JUN, C.-H., 2009. A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications 36*, 2, 3336-3341.

[11] WANG, D., SHEN, Z., SHAO, J., ZHANG, W., XUE, X., and ZHANG, Z., 2016. Multiple Granularity Descriptors for Fine-Grained Categorization. In *IEEE International Conference on Computer Vision*, 2399-2406.

[12] ZHANG, H., XU, T., ELHOSEINY, M., HUANG, X., ZHANG, S., ELGAMMAL, A., and METAXAS, D., 2016. SPDA-CNN: Unifying Semantic Part Detection and Abstraction for Fine-Grained Recognition. In *Computer Vision and Pattern Recognition*, 1143-1152.

[13] ZHANG, N., DONAHUE, J., GIRSHICK, R., and DARRELL, T., 2014. Part-Based R-CNNs for Fine-Grained Category Detection *8689*, 834-849.

[14] ZHANG, X., XIONG, H., ZHOU, W., LIN, W., and TIAN, Q., 2016. Picking deep filter responses for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1134-1142.

[15] ZHANG, Y., WEI, X.-S., WU, J., CAI, J., LU, J., NGUYEN, V.-A., and DO, M.N., 2016. Weakly supervised fine-grained categorization with part-based image representation. *IEEE Transactions on Image Processing 25*, 4, 1713-1725.

[16] ZHANG, Y., WEI, X.S., WU, J., CAI, J., LU, J., NGUYEN, V.A., and DO, M., 2016. Weakly Supervised Fine-Grained Categorization with Part-Based Image Representation. *IEEE Transactions on Image Processing 25*, 4, 1713-1725.

[17] ZHENG, H., FU, J., MEI, T., and LUO, J., 2017. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Int. Conf. on Computer Vision*.