Solution for SnakeCLEF 2022 by Tackling Long-tailed Categorization

Lingfeng Yang[1,2], Xiang Li[3], Renjie Song[2], Kexin Zhu[2] and Gang Li[1]

[1] *Nanjing University of Science and Technology, China*
[2] *Megvii Technology, China*
[3] *Nankai University, China*

Abstract

SnakeCLEF 2022 is a fine-grained image classification benchmark for snake identification. Recently, the masked autoencoder (MAE) has shown superior performance on fine-grained image classification tasks. As a result, we use the MAE pretrained ViT models and refine them on the SnakeCLEF 2022. Overall, the learning process contains two difficulties: 1) dealing with fine-grained species that are visually similar and 2) a long-tailed distribution. To address these issues, we propose using statistic-aware post-processing to process the metadata and refine image predictions. Next, we improve an effective logit adjustment loss (ELAL) to alleviate the classification bias toward the head class. Notably, we achieve 2nd place on the SnakeCLEF 2022 benchmark with a 0.84565 top F1 score. Codes and models are available at https://github.com/ylingfeng/snakeclef2022_fgvc9.

Keywords

SnakeCLEF, Fine-grained image classification, Masked autoencoder, Metadata, Long-tailed distribution

1. Introduction

Fine-grained visual categorization [1, 2, 3, 4, 5, 6, 7] is a popular task to identify fine categories out of coarse divisions. Recently, there is an increasing necessity to develop a fine-grained visual categorization algorithm for various species of snakes for biodiversity, conservation, and global health. The SnakeCLEF 2022 benchmark[1] [8] aims to tackle this requirement, which is held by LifeCLEF [9, 10] jointly with FGVC9[2] of the CVPR 2022.

The difficulty in fine-grained snake identification lies in the high intra-class and low inter-class differences in appearance, and many species are visually similar to others. Moreover, the species distribution in terms of geographical location is irregular, and some countries (e.g., US) contain hundreds of species while some (e.g., Vatican) have only a few types. In addition, the dataset suffers from a severe long-tailed problem in which two-thirds of categories contain less than 100 instances.

In terms of the above problems, we propose to solve them individually. First, as for the visually similar samples which are confusing for image-only predictions, we utilize the metadata [11, 12,

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

yanglfnjust@njust.edu.cn (L. Yang); xiang.li.implus@qq.com (X. Li); songrenjie@megvii.com (R. Song); zhukexin@megvii.com (K. Zhu); gang.li@njust.edu.cn (G. Li)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://www.kaggle.com/competitions/snakeclef2022/

[2] https://sites.google.com/view/fgvc9

13, 14, 15, 16] provided in the dataset to form a prior distribution of whole species. Different from previous multi-modal methods which embed the metadata to the feature space, we design a parameter-free post-processing structure to refine the predictions. To be specific, we record the number of occurrences of metadata corresponding to each species as the priors. More details can be found in Sec. 4.1. Secondly, in Sec. 4.2 we propose the effective logit adjustment loss (ELAL) to alleviate the prediction bias along with training the long-tailed samples by increasing the optimization weight of the tailed classes while reducing the head.

Our contributions can be summarized as:

- We improve a new way to process the metadata by recording statistics referring to each category and a post-processing algorithm is designed to refine the image predictions.
- We propose the effective logit adjustment loss (ELAL) to alleviate the prediction bias resulting from the long-tailed dataset.
- Based on our algorithm, we achieve 2nd place on the SnakeCLEF 2022 benchmark with a 0.84565 top F1 score.

2. Related Work

Fine-grained image classification: To deal with the fine-grained property which is hard to recognize merely through the visual clues, there are three workable solutions: 1) to detect the discriminative regions of an image and pass all parts through the networks for joint classification[1, 3, 4, 6, 7]. 2) Design a robust feature extraction architecture to capture the subtle representations from an image [17, 5, 2, 18, 19]. 3) Utilize the metadata (e.g., shooting date, latitude, longitude, country, and a brief description of the image) [11, 12, 13, 14, 15, 16]. However, the region detector and feature extractor are heavily designed and thus not suitable for our task. Meanwhile, the existing metadata fusion methods all deal with the multimodal feature by embedding them to higher semantic representations before interaction. Specifically, in SnakeCLEF 2022, the types of metadata are discrete (e.g., country, endemic, and code), which is different from the continuous latitude, longitude, or date hypothesized in the previous works. To make use of this metadata, we calculate the existence label within a certain country for all country values in the metadata and form the prior matrix regarding all species.

Long-tailed distribution: In terms of the long-tailed classification, the data re-sampling [20, 21] seeks to change class sampling probability based on the number of samples to get a class-balanced dataset, which includes over-sampling and under-sampling. [22] develop a two-stage paradigm to re-balanced the classifier in the second stage with a frozen backbone. Re-weighting [23, 24, 25, 26] aims to assign the loss weight class-wise to reduce the optimization bias between head-tail classes. The logit adjustment loss [23] encourages a large relative margin between logits of rare versus dominant labels. Based on this work, we modify the margin coefficient and propose an effective logit adjustment loss (ELAL) to solve the long-tailed problem efficiently.

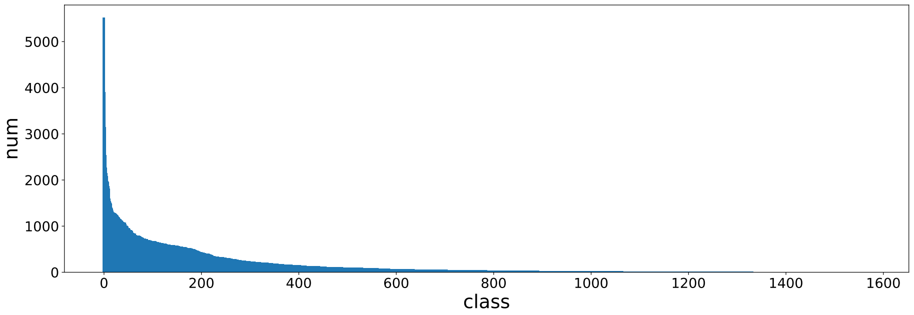

Figure 1: Visualization of the instance number for each class sorted by number in descending order.

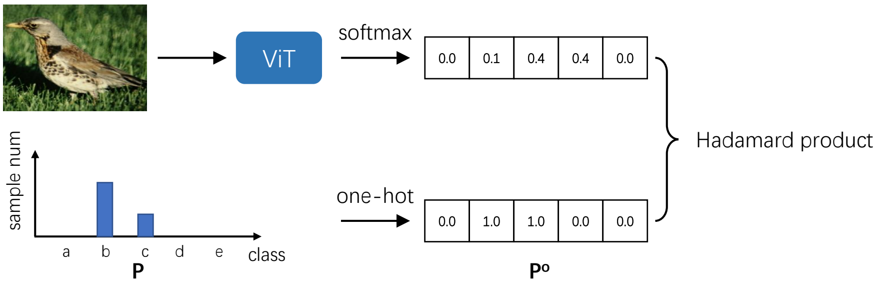

Figure 2: Structure of the post-processing to refine image prediction by category prior extracted from the metadata.

3. Task Description

3.1. Dataset

The SnakeCLEF 2022 dataset [8, 9, 10] is based on observations of 187,129 snakes, containing 318,532 photographs, belonging to 1,572 snake species, observed in 208 countries. The data comes from the online biodiversity platform, iNaturalist. The provided dataset has a heavy long-tailed class distribution (see Fig. 1), where the most frequent species (Natrix natrix) is represented by 6,472 images and the least frequent species by just 5 samples.

3.2. Metric

The evaluation metric for this competition is Mean (Macro) F1-Score. The F1 score, commonly used in information retrieval, measures accuracy using the statistics precision (P) and recall (R). The macro F1 score is not biased by class frequencies and is more suitable for the long-tailed class distributions observed in nature. This metric raises a higher requirement for classification accuracy on tailed categories.

4. Method

4.1. Metadata-aware Post-processing

Given the metadata-label mapping, we count the instance number for all categories if it is attached to a certain metadata value. Then, we obtain the statistic in form of a metadata-wise category matrix $\mathbf{P} \in \mathbb{R}^{n \times c}$, where n is the value number within one type of metadata, and c is the number of classes. Next, we transform $\mathbf{P}$ to a one-hot form $\mathbf{P}^o$, known as the prior statistic, which represents whether a specific category would appear in a certain place. Finally, $\mathbf{P}^o$ is utilized to refine the prediction from the visual networks via the Hadamard product. The whole structure is illustrated in Fig. 2.

4.2. Effective Logit Adjustment Loss

In this section, we introduce our new effective logit adjustment loss (ELAL) function which addresses the performance drop resulting from the prediction bias brought by the long-tailed distribution. First, we give a brief review of the existing loss functions, and then we show how ELAL is developed based on them.

The vanilla softmax cross-entropy can be derived by:

$$\ell(y, f(x)) = \log\left(1 + \sum_{y' \neq y} e^{f_{y'}(x) - f_y(x)}\right), \tag{1}$$

where y denotes the ground-truth label. The logit adjustment loss [23] adds a label-dependent offset to each of the logits, and modifies Eq. 1 with the shifted coefficient M:

$$\ell(y, f(x)) = \log\left(1 + \sum_{y' \neq y} M \cdot e^{f_{y'}(x) - f_y(x)}\right), \tag{2}$$

where $M = \frac{\pi_{y'}}{\pi_y}$, $\pi_y = \frac{N_y}{\sum_{y'} N_{y'}} \in (0, 1)$, and N_y is the total number of instances in each class. Class-balanced Loss [24] proposes the concept of an effective number to replace the direct label-wise instance number to represent the volume of samples. The definition of the effective number is shown as:

$$E_y = \frac{1 - \beta^{N_y}}{1 - \beta}. \tag{3}$$

Inspired by the conception of effective number, which is an improved representation of the vanilla number, we modify the logit adjustment loss by changing the shifted coefficient M to $M = \frac{\epsilon_{y'}}{\epsilon_y}$, $\epsilon_y = \frac{E_y}{\sum_{y'} E_{y'}} \in (0, 1)$ and propose ELAL. Notably, we set $\beta = 1e - 6$ by default in our experiments.

5. Experiments

In this section, we first elaborate on our experimental settings, then ablation studies are conducted to demonstrate the performance of each component. Finally, we list the top results of our methods and give a considerable analysis.

Table 1
Fine-tuning settings on the SnakeCLEF 2022 dataset.

Config	Value
optimizer	AdamW
base learning rate	1e-4 (ViT-L), 1e-3 (ViT-H)
weight decay	0.05
optimizer momentum	$\beta_1, \beta_2{=}0.9, 0.999$
layer-wise lr decay	0.75 (ViT-L), 0.8 (ViT-H)
global batch size (over 8 GPUs)	16
batch size per GPU	2
accumulated iteration	4
learning rate schedule	cosine decay
warmup epochs	5
augmentation	RandAug (9, 0.5)
label smoothing	0.1
mixup	0.8
cutmix	1.0
random erase	0.25
drop path	0.2

5.1. Setup

In this paper, we use the Masked autoencoder (MAE) [27] pretrained ViT [28] models conducted on ImageNet-1K [29] training set for 800 epochs. The fine-tuning codes and checkpoints refer to the MAE repository[3]. The ImageNet-1K dataset has 1.3M images with 1K categories for training and 50K images for validation. Notably, we do not use the larger ImageNet-22K (IN22K) dataset, which contains 14.2M images and 22K classes. Based on the MAE pretrained models, we finetune 50 epochs on the SnakeCLEF 2022 dataset, and the default setting is depicted in Table 1. We randomly select 1/10 of the training dataset to form the validation set to update our algorithm, and a full set is used to train models which present the final submissions. To be specific, we set batch size per GPU to 2 to avoid exceeding the GPU memory. The effective learning rate is obtained following MAE: *lr*= *base_lr*×globalbatchsize / 256. We apply random resizing/cropping, random horizontal flipping [30], label-smoothing regularization [31], Mixup [32], CutMix [33], RandomErasing [34], and RandAug [35] as the standard data augmentations. Notably, all ablation studies are conducted under ViT-L for fair comparisons. The ViT-large and ViT-huge models are trained on eight NVIDIA TITAN Xp GPUs (12G) and eight GeForce RTX 3090 GPUs (24G), respectively.

5.2. Ablation Study

First, we compare the performance with different sets of metadata for post-processing. Table 2 shows that refining predictions with “endemic” and “code” metadata perform the best.

Next, we conduct the ablation on two losses. Table 3 shows our ELAL achieves a higher F1 score under two sets of input resolution. To demonstrate the effectiveness of ELAL on tail

[3] https://github.com/facebookresearch/mae

Table 2
Ablation study on the performance of post-processing under different metadata combinations.

code	endemic	country	val acc	val F1	test F1
			80.470	0.758	0.755
		✓	88.554	0.810	0.796
✓			88.613	0.856	0.834
✓	✓		89.949	0.873	0.864
✓	✓	✓	93.893	0.920	0.815

Table 3
Ablation study on the performance of the long-tailed loss. **CE**: Cross-entropy loss. **ELAL**: Effective logit adjustment loss.

resolution	loss	val acc	val F1	test F1
224	CE	0.858	0.821	0.735
	ELAL	0.915	0.892	0.756
384	CE	0.889	0.859	0.792
	ELAL	0.939	0.920	0.815

Table 4
Ablation study on the performance of the head and tail class. We depict the accuracy of the top 10/50/100/500 from the head/tail classes. **CE**: Cross-entropy loss. **ELAL**: Effective logit adjustment loss.

loss	class	10	50	100	500
CE	head	1.00	1.00	0.95	0.94
	tail	0.30	0.46	0.56	0.79
ELAL	head	1.00	1.00	0.94	0.94
	tail	0.90	0.82	0.88	0.93

Table 5
Performance of the final submissions on public/private benchmarks.

		center crop		multi crop	
model	resolution	public	private	public	private
large	384	0.87134	0.81199	0.87996	0.81997
large	432	0.88375	0.82382	0.89173	0.83063
huge	392	0.89692	0.83662	0.89449	0.84057
ensemble	–	0.90245	0.84409	0.89822	0.84565

class and the potential side effect on head class, we calculate the validation accuracy on the top 10/50/100/500 class from the head and tail classes, respectively (Table 4).

5.3. Results

Based on the strong ViT-L and ViT-H [28], we conduct experiments with an input resolution of 384/392/448 learned on a full training set. We adopt multi-crop [36] as post-processing strategies, which would crop the given image into four corners and the central crop plus the flipped version and average the predictions of whole crops. The model ensemble is an averaging operation over each prediction score after the softmax of the selected models. Our final submissions come from the ensemble of models w/o and w/ multi-crop, which receives 0.84409 and 0.84565 F1 scores on the private benchmark.

5.4. Analysis

We attempt to run a ViT-H with a 448 resolution, which is capable of reaching a higher accuracy theoretically, however, due to the resource limitation we only present the result of a 392 resolution. Also, we notice that the effect of post-processing on the private benchmark is not as significant as on the public benchmark. We suspect that there is a distribution gap between the train and test set of metadata while the public benchmark is less affected.

6. Conclusion

In this paper, we give our solution to the Snake Recognition Competition (SnakeCLEF 2022) in FGVC9, which is challenging due to the fine-grained categorization and long-tailed classes. To deal with the difficulties, we utilize statistic-aware metadata to refile image predictions through post-processing and propose the effective logit adjustment loss (ELAL) to handle the long-tailed problem, respectively. Our team achieves the 2nd place on the private benchmark with a 0.84565 top F1 score.

References

[1] N. Zhang, J. Donahue, R. Girshick, T. Darrell, Part-based r-cnns for fine-grained category detection, in: ECCV, 2014.

[2] Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele, Evaluation of output embeddings for fine-grained image classification, in: CVPR, 2015.

[3] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, Z. Zhang, The application of two-level attention models in deep convolutional neural network for fine-grained image classification, in: CVPR, 2015.

[4] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, L. Wang, Learning to navigate for fine-grained classification, in: ECCV, 2018.

[5] D. Chang, Y. Ding, J. Xie, A. K. Bhunia, X. Li, Z. Ma, M. Wu, J. Guo, Y.-Z. Song, The devil is in the channels: Mutual-channel loss for fine-grained image classification, IEEE TIP (2020).

[6] F. Zhang, M. Li, G. Zhai, Y. Liu, Multi-branch and multi-scale attention learning for fine-grained visual categorization, in: MMM, 2021.

[7] A. Behera, Z. Wharton, P. Hewage, A. Bera, Context-aware attentional pooling (cap) for fine-grained visual classification, arXiv preprint arXiv:2101.06635 (2021).

[8] L. Picek, A. M. Durso, M. Hrúz, I. Bolon, Overview of SnakeCLEF 2022: Automated snake species identification on a global scale, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.

[9] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, I. Bolon, et al., Lifeclef 2022 teaser: An evaluation of machine-learning based species identification and species distribution prediction, in: European Conference on Information Retrieval, Springer, 2022, pp. 390–399.

[10] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, H. Glotin, R. Planqué, W.-P. Vellinga, A. Navine, H. Klinck, T. Denton, I. Eggel, P. Bonnet, M. Šulc, M. Hruz, Overview of lifeclef 2022: an evaluation of machine-learning based species identification and species distribution prediction, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2022.

[11] K. Tang, M. Paluri, L. Fei-Fei, R. Fergus, L. Bourdev, Improving image classification with location context, in: ICCV, 2015.

[12] H. C. Wittich, M. Seeland, J. Wäldchen, M. Rzanny, P. Mäder, Recommending plant taxa for supporting on-site species identification, BMC bioinformatics (2018).

[13] G. Chu, B. Potetz, W. Wang, A. Howard, Y. Song, F. Brucher, T. Leung, H. Adam, Geo-aware networks for fine-grained recognition, in: ICCV, 2019.

[14] O. Mac Aodha, E. Cole, P. Perona, Presence-only geographical priors for fine-grained image classification, in: ICCV, 2019.

[15] L. Yang, X. Li, R. Song, B. Zhao, J. Tao, S. Zhou, J. Liang, J. Yang, Dynamic mlp for fine-grained image classification by leveraging geographical and temporal information, arXiv preprint arXiv:2203.03253 (2022).

[16] Q. Diao, Y. Jiang, B. Wen, J. Sun, Z. Yuan, Metaformer: A unified meta framework for fine-grained recognition, arXiv preprint arXiv:2203.02751 (2022).

[17] H. Zheng, J. Fu, Z.-J. Zha, J. Luo, Learning deep bilinear transformation for fine-grained image representation, arXiv preprint arXiv:1911.03621 (2019).

[18] Y. Gao, X. Han, X. Wang, W. Huang, M. Scott, Channel interaction networks for fine-grained image categorization, in: AAAI, 2020.

[19] H. Touvron, A. Sablayrolles, M. Douze, M. Cord, H. Jégou, Grafit: Learning fine-grained image representations with coarse labels, in: ICCV, 2021.

[20] B. Zhou, Q. Cui, X.-S. Wei, Z.-M. Chen, Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition, in: CVPR, 2020.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research (2002).

[22] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, Y. Kalantidis, Decoupling representation and classifier for long-tailed recognition, arXiv preprint arXiv:1910.09217 (2019).

[23] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, S. Kumar, Long-tail learning via logit adjustment, arXiv preprint arXiv:2007.07314 (2020).

[24] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: CVPR, 2019.

[25] J. Tan, X. Lu, G. Zhang, C. Yin, Q. Li, Equalization loss v2: A new gradient balance approach for long-tailed object detection, in: CVPR, 2021.

[26] B. Li, Y. Yao, J. Tan, G. Zhang, F. Yu, J. Lu, Y. Luo, Equalized focal loss for dense long-tailed object detection, arXiv preprint arXiv:2201.02593 (2022).

[27] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision

learners, in: CVPR, 2021.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: CVPR, 2009.

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: CVPR, 2015.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: CVPR, 2016.

[32] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, arXiv preprint arXiv:1710.09412 (2017).

[33] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: ICCV, 2019.

[34] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: AAAI, 2020.

[35] E. D. Cubuk, B. Zoph, J. Shlens, Q. V. Le, Randaugment: Practical automated data augmentation with a reduced search space, in: CVPRW, 2020.

[36] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: NeurIPS, 2012.