A Deep Learning based Solution to FungiCLEF2023

Feiran Hu[1], Peng Wang[1], Yangyang Li[1], Chenlong Duan[1], Zijian Zhu[1], Yong Li[1,*] and Xiu-Shen Wei[1,*]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China

Abstract

The FungiCLEF2023 competition intends to foster the development of advanced algorithms for fungi species identification through the analysis of images and metadata, thereby making notable contributions to biodiversity conservation and human health. To overcome the inherent challenges posed by this competition, this paper introduces a deep learning based approach to solving this problem, which is based on the VOLO [1] backbone architecture with rich data augmentation and the loss function addressing data imbalance issues. We also discuss methods for utilizing metadata to address the issue of open-set recognition under limited model capacity. Our method achieves 54.34% on private leaderboard, which is the third place among the participators. The code is available at https://github.com/xiaoxsparraw/CLEF2023.

Keywords

Fungi Species Identification, Fine-grained image recognition, Open-Set, Long-tailed, Metadata

1. Introduction

Fine-grained visual categorization represents a fundamental and significant challenge in computer vision and pattern recognition, playing a crucial role in various practical applications [2]. The FungiCLEF2023 competition, held jointly as part of the LifeCLEF2023 lab in the CLEF2023 conference and the FGVC10 workshop organized in conjunction with the CVPR2023 conference, aims to advance the progress of robust algorithms for the identification of fungi species using image and metadata inputs. Achieving this objective carries profound implications for biodiversity conservation and plays a vital role in preserving human health.

Previous iterations of this competition have achieved remarkable performances accomplished by deep learning models [3, 4, 5, 6, 7]. In order to augment the practical significance of the competition and effectively address the concerns encountered by developers, scientists, users, and communities, the organizers introduce more constraints. Consequently, the challenges encountered in this year's competition can be summarized as follows:

- **Fine-grained image recognition:** Fine-grained image analysis has long presented a challenge within the FGVC workshop, prompting the need for further investigation and research.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.

hufr@njust.edu.cn (F. Hu); wangpeng@njust.edu.cn (P. Wang); lyylyyi599@njust.edu.cn (Y. Li); duancl@njust.edu.cn (C. Duan); zhuzj@njust.edu.cn (Z. Zhu); yong.li@njust.edu.cn (Y. Li); weixs.gm@gmail.com (X.-S. Wei)

- **Open-Set recognition:** Open-Set recognition constitutes a critical problem to be addressed in real-world applications. The FungiCLEF challenge's test dataset includes numerous species that do not appear in the training dataset.
- **Utilization of metadata:** The incorporation of metadata assumes a vital role in the classification process, particularly in the identification of fungi species. Such metadata, commonly utilized by individuals in their everyday lives, necessitates utilization of location-based information in deep learning models.
- **Long-tailed distribution:** The prevalence of long-tailed distribution permeates numerous real-world scenarios, including the distribution of fungi species. Managing the challenges posed by the long-tailed nature of fungi species distribution calls for corresponding solutions.
- **Model size limitation:** A strict constraint has been imposed on the model size, limiting it to a maximum of 1GB.

The challenges mentioned above pose significant difficulties that require to overcome. In light of this, this paper proposes a preliminary solution utilizing deep learning techniques. Our method is based on VOLO [1], strong data augmentation techniques are utilized by us too. Besides, seesaw loss [8] helps to relieve the challenge of long-tailed distribution. To recognize the open-set classes in test dataset, a post-processing method is taken by us, which means assign -1 prediction to examples that the model is uncertain about.

The subsequent sections of this paper offer overview of the key components. Section 2 provides a detailed explanation of the datasets, accompanied by examination of the evaluation metric employed. In Section 3, we discuss our proposed methodologies in detail. Section 4 presents the implementation details, coupled with analysis of the results and findings obtained. Lastly, in Section 5, we conclude this paper by summarizing the key insights derived from the study and discussing potential directions for future research.

2. Dataset and Evaluation Metric

Understanding of datasets and metrics is a fundamental requirement for effectively participating in a competition. In this section, we show our comprehension of the datasets utilized in the competition and provide an overview of the evaluation metrics selected by the competition organizers.

2.1. Dataset

The challenge dataset primarily relies on the data sourced from the Danish Fungi 2020 dataset [9], comprising 295,938 training images, each corresponding to one of the 1,604 observed species, predominantly found within Denmark. All training samples passed an expert validation process, ensuring high-quality labels. Additionally, the dataset includes observation metadata encompassing details about habitat, substrate, time, location, etc. In conjunction with the training dataset, the organizers have provided a validation set. This set comprises 30,131 observations, encompassing 60,832 images, and spanning across 2,713 species. Importantly, this validation set includes observations collected throughout the entire year, encompassing diverse substrate

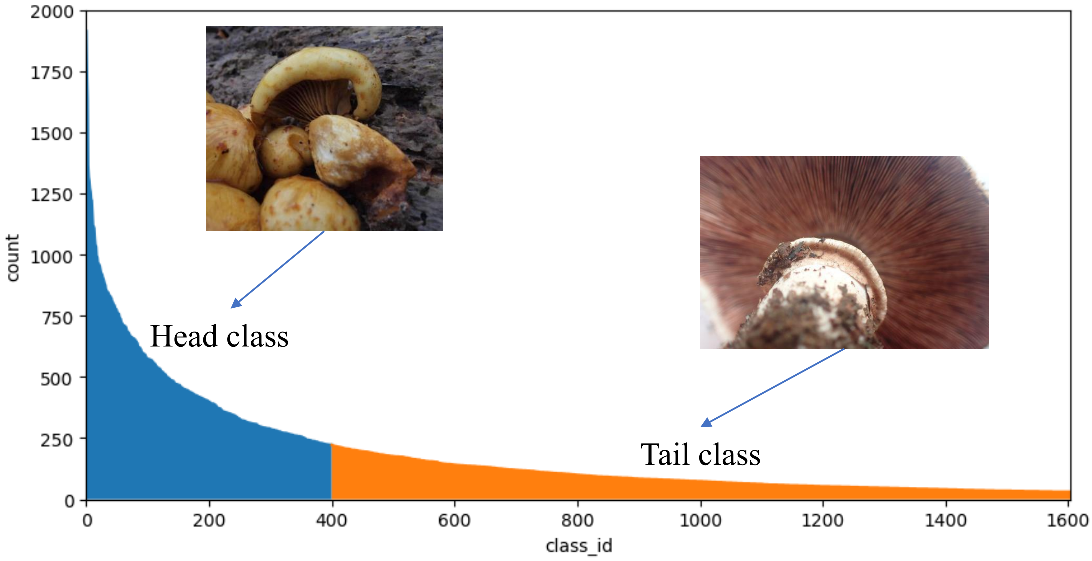

Figure 1: Long-tailed distribution of the FungiCLEF2023 training dataset. The classes are arranged in descending order based on the number of samples in each category, with the color of blue representing the head class and the orange representing the tail class.

and habitat types. Furthermore, a public test set has been made available, consisting of 30,130 observations and 60,225 corresponding images. The data distribution within the public test set closely aligns with that of the validation set, ensuring consistency and facilitating a fair evaluation of the models' performance.

It is important to acknowledge the presence of a significantly imbalanced long-tailed distribution within the given dataset, as illustrated in Figure 1. This distribution is characterized by a substantial disparity in the number of instances across different classes, with a few classes having a disproportionately large number of samples compared to the majority of classes. Furthermore, it should be noted that the test set includes multiple out-of-scope classes, which pose an additional challenge in the evaluation process.

2.2. Evaluation Metric

The FungiCLEF2023 recognition task incorporates several evaluation metrics, as shown by Equation 1, which cater to distinct decision problems. The primary objective is to minimize the empirical loss $\mathbf{L}$ for decision $q(x)$ across a set of observations x and their corresponding true labels y. This optimization process is achieved by considering a cost function $W(y, q(x))$ that quantifies the impact of the decision $q(x)$ on the true label y:

$$\mathbf{L} = \sum_{i=1}^{5} W(k_i, q(x_i)) \, . \tag{1}$$

The first metric is the standard classification error. All species not represented in the training set should correctly be classified as an "unknown" category. The second one is the cost for confusing edible species for poisonous and vice versa. The third one is a user-focused loss composes of both the classification error and the poisonous confusion. The fourth one is the cost for confusing “unknown” species, that missing “unknown” species is higher than misclassifying species. The last one is an increasing the weight of performance on rare species. While improvements in average classification accuracy can be improved by focusing on the more common species, in some applications correct classification of rare species is of high importance. We set the misclassification cost proportional to the inverse frequency of the species in the training dataset. Note that if the species distribution was the same in the test set, this would be equivalent to macro-averaged accuracy; but it is not the case of the FungiCLEF2023 test set.

3. Method

In this section, we will introduce our solution to FungiCLEF2023.

3.1. Data Augmentation

Data augmentation plays a crucial role in computer vision tasks, and it is an integral part of our methodology. In our method, we employ a set of fundamental image augmentation methods sourced from Albumentations [10]. These methods encompass a range of transformations, including RandomResizedCrop, Transpose, HorizontalFlip, VerticalFlip, ShiftScaleRotate, RandomBrightnessContrast, PiecewiseAffine, HueSaturationValue, OpticalDistortion, ElasticTransform, Cutout, and GridDistortion. In addition to these standard augmentation techniques, we also incorporate data mixing augmentation methods, namely Mixup [11], CutMix [12], TokenMix [13], and RandomMix [14], during the course of the competition. These data mixing techniques provide effective regularization for the models by blending both images and labels, thereby mitigating the risk of overfitting to the training dataset. This approach helps to enhance the generalization capability of the models and improve their performance on unseen data.

3.2. Backbone

Throughout the competition, we conducted exploration of various models, including both classical and state-of-the-art architectures, to develop robust solutions. These models encompassed Convolutional Neural Networks and Vision Transformers, which have demonstrated remarkable performance in computer vision tasks. Notable models employed during the competition include ResNet [15], VOLO [1], BEiT-v2 [16], and ConvNeXt-v2 [17]. The implementation of these models was facilitated using the timm library [18], which offers a versatile framework for model development and evaluation. Through extensive experimentation under various settings, we selected VOLO [1] as the backbone architecture for our final proposed method. This decision was based on the model’s superior performance and its ability to effectively capture relevant visual features and patterns for fungi species recognition.

3.3. Optimization Procedure

Dealing with long-tailed recognition constitutes a significant challenge encountered in the competition. To address this issue, we conducted a thorough investigation of various techniques, leveraging the insights and methods proposed in BagofTricks-LT [19]. In our final submission, we strategically integrated the seesaw loss [8] as a fundamental component of our approach. The seesaw loss formulation is expressed as follows:

$$\begin{aligned} L_{\text{seesaw}}(\mathbf{z}) &= -\sum_{i=1}^{C} y_i \log\left(\widehat{\sigma}_i\right), \\ \text{with } \widehat{\sigma}_i &= \frac{e^{z_i}}{\sum_{j \neq i}^{C} \mathcal{S}_{ij} e^{z_j} + e^{z_i}}, \end{aligned} \tag{2}$$

where $\mathbf{z}$ refers to the output obtained from the fully connected layer, C represents the total number of classes, and y_i denotes the one-hot label of the image under consideration. The hyper-parameters $\mathcal{S}_{ij}$ are determined, taking into account the distribution characteristics inherent in the dataset. These hyper-parameters play a crucial role in guiding the training process and ensuring the appropriate handling of class imbalance during optimization.

In addition to the careful selection of loss functions, the choice of an optimizer and an appropriate learning rate decay strategy are crucial factors in the training of our models. For optimization, we utilize the AdamW [20] optimizer. To further enhance the convergence speed and overall performance, we incorporate the cosine learning rate decay technique [21] in conjunction with warmup techniques during the training process. This combination of strategies promotes more efficient and effective model convergence, leading to improved training outcomes.

3.4. Post-processing

In the context of open-set setting, accurately determining whether a test sample belongs to a known category is important for the deployment of a model. Vaze et al. [22] have shown that the classifier's ability to make a "none-of-above" decision is closely linked to its accuracy on closed-set classes. They achieved state-of-the-art performance on existing open-set recognition (OSR) benchmarks by employing a more robust baseline model. However, in this year's challenge, due to limitations imposed by the model and training data, it has been challenging to train a baseline model in closed-set scenarios. Consequently, our focus has shifted towards exploring the utilization of metadata shown in Tabel 1, as an alternative approach.

Inspired by Aodha et al. [23], we developed a prior model trained on a balanced training set, which we refer to as a Multi-Layer Perceptron (MLP), comprising three fully connected layers and employing dropout regularization. We tried to use the metadata of "countryCode", "level2Name", "Latitude" with "Longitude" and "Habitat" with "MetaSubstrate". For text and number, we utilized the text encoder of CLIP [24] and $[sin(\pi x), cos(\pi x)]$ to map, respectively. However, our attempts revealed that utilizing metadata did not yield improvements in the final outcomes. We attribute this to the insufficiency of the baseline model's robustness and its limited resistance to perturbations. After threshold processing, it still led to confusion

Table 1
Description of the provided metadata.

Tags	Description
CountryCode	Country information.
Latitude, Longitude	The location information of latitude and longitude.
Level2Name	The position information of the observer.
Habitat	Observations of the environment.
Substrate	Natural substance on which to live, such as bark, soil, etc.

between unknown and known classes. Therefore, we only use a simple threshold for this year competition.

4. Experiments

In this section, we will introduce our implementation details and main results.

4.1. Experiment Settings

The proposed methodology has been developed using the PyTorch framework [25]. All models utilized in our approach have pre-training on the ImageNet dataset [26], which is readily available within the timm library [18]. Fine-tuning of these models was performed using 4 Nvidia RTX3090 GPUs. The initial learning rate was set to 2×10^{-5}, and the total number of training epochs was set to 15, with the first epoch dedicated to warm-up by employing a learning rate of 2×10^{-7}. For optimal model training, we employed the AdamW optimizer [27] in conjunction with a cosine learning rate scheduler [21], with the weight decay set to 2×10^{-5}. During inference on the test dataset, test time augmentation was incorporated. Additionally, considering that an observation may consist of multiple images, we adopted a simple averaging approach to obtain a single prediction for each observation.

4.2. Results

In this section, we present the main findings obtained during the challenge, as depicted in Table 2. The column labeled "Metric" in the table represents the F1 score on the leaderboard.

Due to the substantial size of the FungiCLEF dataset and inherent constraints in terms of our energy and hardware, our experimentation on FungiCLEF was limited. Furthermore, certain techniques [28] that demonstrated effectiveness in SnakeCLEF during our participation were not employed in FungiCLEF. Initially, we began with the implementation of ResNet [15], and the outcomes indicated that weight cross entropy loss did not contribute significantly to the task. In our preliminary experiments, the inclusion of metadata yielded performance improvements on the leaderboard. However, in more robust models, the incorporation of metadata failed to deliver performance enhancements.

We also incorporated the utilization of Seesaw loss [8] and CutMix [12] in the FungiCLEF task, which proved to be effective. These techniques alleviated the challenges posed by the

Table 2
Results of FungiCLEF.

Backbone	Resolution	Metric (%)	Comments
ResNet50 [15]	224×224	43.19	CE loss
ResNet50 [15]	224×224	40.64	weight CE loss
BEiT-v2-B [16]	224×224	50.40	stronger backbone
BEiT-v2-B [16]	224×224	51.49	metadata
VOLO [1]	448×448	52.65	seesaw loss
VOLO [1]	448×448	53.93	seesaw loss + cutmix
VOLO [1]	448×448	55.46	seesaw loss + cutmix + open-set
ConvNeXt-v2-L [17]	512×512	55.35	seesaw loss + cutmix + open-set

long-tailed distribution and enhanced the generative capability of the models. As shown in Table 2, we observed that as the resolution increased, the metric score plateaued when reaching a resolution of 448. Based on our experiments detailed in Table 2, our final submission employed VOLO [1] as the backbone model, without the utilization of any metadata.

5. Conclusion

Fine-grained visual analysis continues to present significant challenges, particularly in the domain of fungi species recognition, where the shared visual characteristics among different species, coupled with their prevalence in humid environments, contribute to the complexity of the task. In our approach, we solely rely on image data for predictions, disregarding the available metadata entirely. The open-set problem and the integration of metadata and visual data present a persistent challenge that demands further research and exploration.

References

[1] L. Yuan, Q. Hou, Z. Jiang, J. Feng, S. Yan, Volo: Vision outlooker for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

[2] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, S. Belongie, Fine-grained image analysis with deep learning: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (2021) 8927–8948.

[3] L. Picek, M. Šulc, J. Matas, J. Heilmann-Clausen, Overview of fungiclef 2022: Fungi recognition as an open set classification problem, Working Notes of CLEF (2022).

[4] J. Yu, H. Chang, K. Lu, G. Xie, L. Zhang, Z. Cai, S. Du, Z. Wei, Z. Liu, F. Gao, et al., Bag of tricks and a strong baseline for FGVC, Working Notes of CLEF (2022).

[5] G. Fan, C. Zining, W. Weiqiu, S. Yinan, S. Fei, Z. Zhicheng, C. Hong, Does closed-set training generalize to open-set recognition?, Working Notes of CLEF (2022).

[6] K. Desingu, A. Bhaskar, M. Palaniappan, E. A. Chodisetty, H. Bharathi, Classification of fungi species: A deep learning based image feature extraction and gradient boosting ensemble approach, Working Notes of CLEF (2022).

[7] S. Wolf, J. Beyerer, Transformer-based fine-grained fungi classification in an open-set scenario, Working Notes of CLEF (2022).

[8] J. Wang, W. Zhang, Y. Zang, Y. Cao, J. Pang, T. Gong, K. Chen, Z. Liu, C. C. Loy, D. Lin, Seesaw loss for long-tailed instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9695–9704.

[9] L. Picek, M. Šulc, J. Matas, T. S. Jeppesen, J. Heilmann-Clausen, T. Læssøe, T. Frøslev, Danish fungi 2020-not just another image recognition dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1525–1535.

[10] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A. A. Kalinin, Albumentations: Fast and flexible image augmentations, Information 11 (2020) 125.

[11] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, in: International Conference on Learning Representations, 2018.

[12] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6023–6032.

[13] J. Liu, B. Liu, H. Zhou, H. Li, Y. Liu, Tokenmix: Rethinking image mixing for data augmentation in vision transformers, in: European Conference on Computer Vision, Springer, 2022, pp. 455–471.

[14] X. Liu, F. Shen, J. Zhao, C. Nie, Randommix: A mixed sample data augmentation method with multiple mixed modes, arXiv preprint arXiv:2205.08728 (2022).

[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[16] Z. Peng, L. Dong, H. Bao, Q. Ye, F. Wei, Beit v2: Masked image modeling with vector-quantized visual tokenizers, arXiv preprint arXiv:2208.06366 (2022).

[17] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, S. Xie, Convnext v2: Co-designing and scaling convnets with masked autoencoders, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 16133–16142.

[18] R. Wightman, Pytorch image models, https://github.com/rwightman/pytorch-image-models, 2019.

[19] Y. Zhang, X. Wei, B. Zhou, J. Wu, Bag of tricks for long-tailed visual recognition with deep convolutional neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 3447–3455.

[20] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019.

[21] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, in: International Conference on Learning Representations, 2017.

[22] S. Vaze, K. Han, A. Vedaldi, A. Zisserman, Open-set recognition: A good closed-set classifier is all you need, in: International Conference on Learning Representations, 2022.

[23] O. Mac Aodha, E. Cole, P. Perona, Presence-only geographical priors for fine-grained image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9596–9606.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language

supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[27] I. Loshchilov, F. Hutter, Fixing weight decay regularization in adam (2017).

[28] F. Hu, P. Wang, Y. Li, C. Duan, Z. Zhu, F. Wang, F. Zhang, Y. Li, X.-S. Wei, Watch out venomous snake species: A solution to snakeclef2023, in: CLEF 2023-Conference and Labs of the Evaluation Forum, 2023.