

Does Closed-Set Training Generalize to Open-Set Recognition?

Fan Gao¹, Zining Chen¹, Weiqiu Wang¹, Yinan Song¹, Fei Su¹, Zhicheng Zhao¹ and Hong Chen²

¹Beijing Key Laboratory of Network System and Network Culture, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

²China Mobile Research Institute

Abstract

Automatic classification of fungi assists scientists in species identification and biodiversity protection. The FungiCLEF 2022 challenge provides a large-scale multi-modal fine-grained dataset to contribute to this issue. This paper proposes a novel open-set image classification method called Class-wise Weighted Prototype Classifier (CWPC) which decouples closed-set training and open-set inference. Thus, it can benefit from all existing closed-set advances and transfer to open-set without further modification. By using meta-vision models and two different vision-only models, an ensemble result achieves excellent performance with the mean F1 scores of 81.02% and 77.58% on public leaderboard and private leaderboard, respectively.

Keywords

Fungi identification, Fine-grained, Open-set recognition, Metadata, Long-tailed

1. Introduction

Fungi contains many fine-grained classes of eukaryotic organisms that are widely distributed in nature and play an important role in human production and life. Automatic recognition of fungi species assists mycologists, citizen scientists and nature enthusiasts in species identification in the wild. However, fungi identification is difficult because of the high diversity of fungi, fine granularity of species and domain gap caused by observation tools. As a part of LifeCLEF-2022 [1, 2] which aims at biodiversity identification and prediction, FungiCLEF-2022 [3] searches for a robust open-set fungi identification system which is more practical than a closed-set recognition system in real-world scenario.

Thanks to the large-scale labeled dataset such as ImageNet [4] and iNaturalist [5], convolution neural networks are the mainstream of vision recognition and outperform human experts in some fields [4, 6]. Due to the limitation of convolution structures, CNN only takes image information as input and can't benefit from rich metadata information. Meanwhile, most of existing methods and optimization are based on closed-set recognition which can't be applied to open-set recognition directly.

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

✉ xsgaofan@bupt.edu.cn (F. Gao); chenzn@bupt.edu.cn (Z. Chen); wangweiqiu@bupt.edu.cn (W. Wang);

songyn@bupt.edu.cn (Y. Song); sufei@bupt.edu.cn (F. Su); zhaozc@bupt.edu.cn (Z. Zhao);

chenhongyj@chinamobile.com (H. Chen)

© 2022 Copyright for this paper by its authors. <https://www.overleaf.com/project/62b09d47feab3260d8e84f90> Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



CEUR Workshop Proceedings (CEUR-WS.org)

In this paper, we propose a novel open-set classification method called Class-wise Weighted Prototype Classifier (CWPC) by decoupling closed-set training and open-set inference. On the one hand, the open-set recognition can benefit from all the advances in closed-set image classification, such as large-scale pre-trained models, label smoothing and data augmentation. On the other hand, it is cost-saving to transfer closed-set recognition models to open-set scenarios with further modification and training. As for the closed-set training, firstly we elaborately design a text template to compensate the context of metadata which have more reasonable and complete semantic information than discrete and independent words. And then, we combine text and vision information with a meta-vision model where convolution is used to extra deep vision embedding and transformer is used to fuse image and metadata information. We also employ two different vision-only models to complement each other. A hard classes mining strategy and LDAM loss [7] are used to eliminate the long-tailed distribution of dataset. Finally, we got a result of 81.02% and 77.58% with our method respectively on public leaderboard and private leaderboard.

2. Related Work

2.1. Open-set Recognition

Unlike traditional closed-set recognition, open-set recognition is more suitable for real-world applications. This task was first proposed in [8], in which the authors apply an 1-vs-Set machine to calculate an open-space risk as an indicator. When a sample is far from known samples, the increased risk suggests it is more likely from unknown classes. OpenMax [9] replaces the SoftMax layer in DNNs with an OpenMax layer to redistribute to get the class probability of unknown samples. PROSER [10] takes open-set problem into consideration during the training process. It generates data placeholders by fusing middle hidden layer features from different classes as the embedding of open-set classes and augments the output layer with an extra dummy classifier to well separate known and unknown. Although these methods make great progress on open-set recognition, they can't utilize metadata efficiently.

2.2. Multi-Modality

Fine-grained classification methods with only images have been explored by many researchers. Besides visual information, additional information is used to improve the performance. CVL [11] proposes a two-branch network while the vision stream learns deep vision representations and the language stream learns text representations. The results of two streams are merged in later stage to combine vision and language. Geo-Aware Networks [12] incorporates geolocation information prior to fine-grained classification and examines various ways of geographic prior. MetaFormer [13] is a hybrid structure backbone where the convolution can extra image embedding and introduce the inductive bias of the convolution, and the transformer can fuse visual and meta-information.

2.3. Long-tail Recognition

For long-tailed recognition, re-balancing methods including re-weighting [14, 15] and re-sampling [16, 17] are conventional methods to alleviate the imbalance of datasets. However, recent studies find that they may do harm to feature learning. Besides, re-balancing methods are easily over-fitting and under-fitting on the tail and head classes, respectively. Multi-experts models such as BBN [18] and RIDE [19], are also designed to solve long-tailed problem, but these methods have high computation complexity and are hard to optimize when we choose a large-scale pretrained model as backbone. OLTR [20] is the first work proposed for open-set long-tailed recognition which utilizes extra attention module and memory bank. Therefore, considering the computation and memory cost, in our strategy, we apply LDAM loss on the large-scale pretrained models to fine-tune on the competition dataset, which assigns large margins on the high-frequency classes and small margins on the low-frequency ones.

3. Challenge Description

3.1. Dataset

The data of FungiCLEF is from Danish Fungi 2020 [21], a novel fine-grained dataset which consists of 266,344 images for training and 29,594 images for validation. It contains 1,604 species mainly from the Fungi kingdom with a few visually similar species. While most of images are collected from natural scene, there remains some hand drawn drafts and microscope observations which have a huge domain gap with others, as shown in Figure 1. In addition to image information and class labels, this dataset provides rich observation metadata in csv files. There are more than 20 kinds covering basic time and geographic localities, full taxonomy labels, substrate and habitat, etc. For some images, not all meta information is available and some are missing. The class frequencies in the dataset follow an extremely unbalanced long-tailed distribution with a maximum 1,913 and a minimum 31, as illustrated in Figure 2. An additional set of 118,676 images from 3,134 species is used for testing. These images are provided with less metadata (e.g. time stamp, location, substrate, habitat).

3.2. Task

Being a part of LifeCLEF-2022 which aims at biodiversity identification and prediction, FungiCLEF-2022 is an automatic fungi recognition competition, as well as an open-set machine learning problem, which means unknown categories will emerge during test time. Under this circumstances, open-set recognition task is proposed to perform on known classes and reject unknown classes as one class. Meanwhile, Danish Fungi 2020 is a fine-grained dataset with 1,604 fungi species. Small inter-class variances and huge intra-class similarity make it more challenging. Contrast to traditional visual recognition, this task provides rich metadata acquired by citizen-scientists, i.e. only vision models are not sufficient, the combination of metadata and images must be considered. Here we conclude the main difficulties of this completion:

- Usage of rich metadata;
- Extremely unbalanced long-tailed data distribution;

- Open-set recognition rather than closed-set;
- Robust recognition with noise data, e.g. images, hand-drawn drafts and microscopic observations.

4. Method

In this section, we will introduce our solution for the open-set fungi recognition challenge. The insight of our solution is to generalize models trained on the closed-set dataset to the open-set scenario without any additional trivial module or extra computation cost on open-set training. Therefore, we decouple the open-set recognition into closed-set training and open-set inference, described in Section 4.1 and Section 4.2, respectively. For closed-set training, we utilize the existing closed-set advances and innovate to use metadata with a designed text template and merge multi-modal embeddings in feature space. For open-set inference, we design a Class-wise Weighted Prototype Classifier (CWPC) and the ObservationId-awared Weighted Similarity (OAWS) strategy to generalize closed-set training models to open-set recognition challenge. Besides, we proposed a weighted Top-5 voting strategy to ensemble diverse models for better performances.

4.1. Closed-Set Training Improvements

4.1.1. Multi-modal Information Usage

Metadata Preprocessing. For training and validation data, more than 20 kinds of metadata are provided including time stamp, geographic localities, full taxonomy labels, substrate and habitat, etc. There are plenty of choices during training, while it only provides 10 metadata

Figure 1: Samples of fungi challenge dataset.

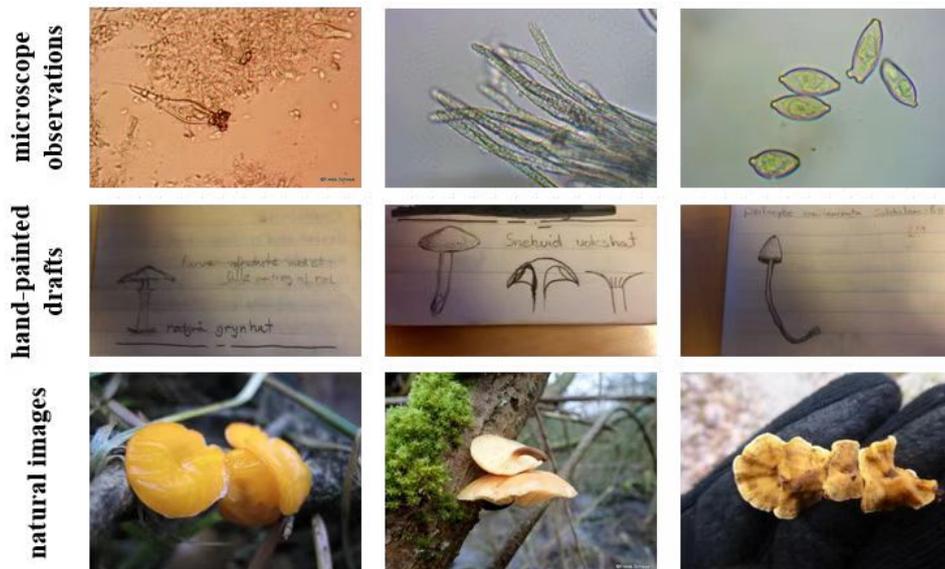
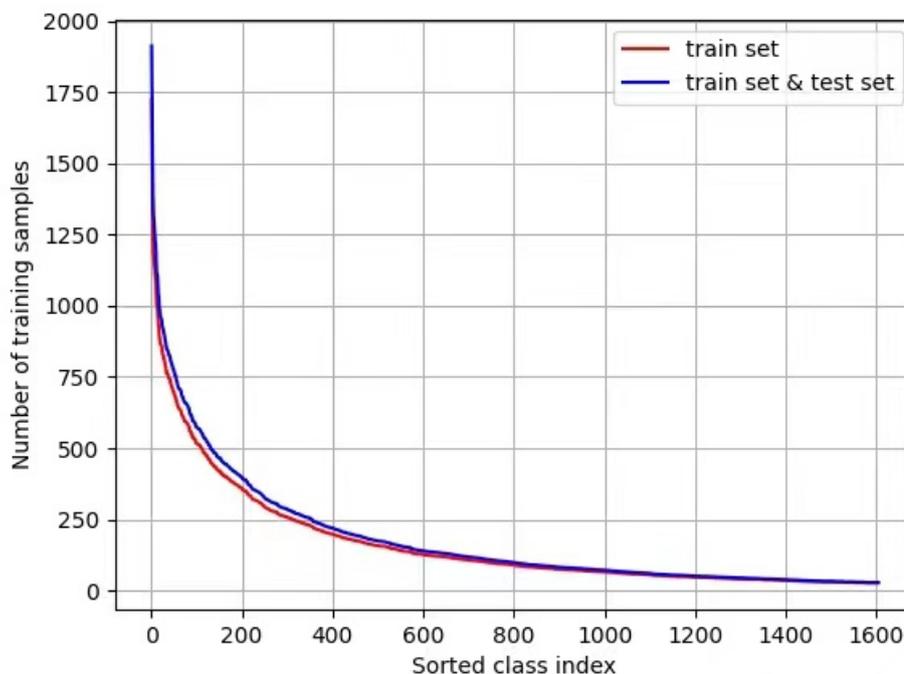


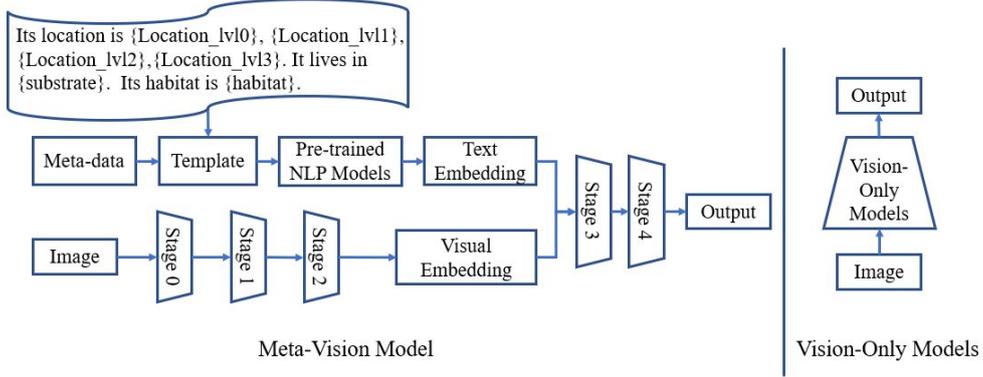
Figure 2: Distribution of fungi challenge dataset.



for test set: "eventDate", "month", "day", "countryCode", "Location_lvl0", "Location_lvl3", "Location_lvl2", "Location_lvl1", "Substrate" and "Habitat". Therefore, to keep the consistency of training and testing, we choose from the above ten metadata dropping time-related ones in consideration of the potential confusion caused by time. Also we replace "countryCode" with full country name. Instead of regarding these metadata as discrete and independent words, we design a description text template with all metadata and replace the missing ones with the word "unknown". For example, if the values of "countryCode", "Location_lvl0", "Location_lvl3", "Location_lvl2", "Location_lvl1", "Substrate" and "Habitat" are "US", "Mount Olive Baptist Church", "United States", "Texas", "Brazoria", "bark of living trees" and "None", respectively, with our template, the description is: "Its location is Mount Olive Baptist Church, Brazoria, Texas, United States. It lives in bark of living trees. Its habitat is unknown". And the description is taken as the caption of its corresponding image, which is used in Metadata Encoding. Our designed text template eliminates the distractions of missing metadata, adds contextual information and ensures that we can get fixed dimension features for later stages.

Metadata Encoding. To get deep text embedding efficiently, we employ pre-trained NLP models directly. Intuitively, we use a multilingual BERT [22]-base model because the location is recorded in Danish. It is pretrained on the top 104 languages with the largest Wikipedia using masked language modeling (MLM) objective. And for each designed template, it generates a 768-dimension feature. Further, we update the multilingual BERT model with RoBERTa [23]

Figure 3: The pipeline of Meta-vision Models and vision-only models during closed-set training.



large model. RoBERTa is a well-trained BERT with some modifications including more epochs, larger batches, more data, etc. It generates a 1,024-dimension feature for each text template which contains more information and can be more representative.

Meta-Vision Models. We use MetaFormer as our meta-vision backbone to add meta information to improve the fine-grained classification. Metaformer is a hybrid framework which uses convolution to extract deep vision features and uses transformer layers to fuse vision and meta information. The origin MetaFormer design multi-layered fully-connected networks for each metadata to get embedding vector. However, our meta information has been merged as one in a unified text template and encoded by pre-trained NLP models as described in Metadata Preprocessing and Metadata Encoding, respectively. After getting initial text embeddings with a pre-trained NLP model, we apply a single fully-connected layer on them, followed by an activation layer ReLU and layer normalization. Relative transformer layers in MetaFormer are used to fuse visual token, meta token and class token. Like ViT [24], only the class token is used for the category prediction.

Vision-Only Models. While MetaFormer focuses on the fusion of multi-modal information, models only trained on images is of necessity to learn visual-representative deep features. Here we use convolution-based ConvNeXt [25] and transformer-based Swin Transformer [6], both of which are pioneer works in their respective fields. We hope these two different network structures can pay attention to different image patterns, bring new views into learning process and complement each other in the final decision. We adopt vanilla Swin Transformer and ConvNeXt architecture for simplicity.

To sum up, during closed-set training, the pipeline of meta-vision models and vision-only models are illustrated in Figure 3.

4.1.2. Long-Tailed Solution

LDAM Loss [7]. As analyzed in Section 3.1, the dataset shows an extremely unbalanced long-tailed distribution, which will deteriorate the network performance during testing. To alleviate this adverse effect, we train our models with LDAM loss rather than CE loss. LDAM loss enforces a class-aware margin for each class to optimize a uniform-label generalization

Figure 4: Dirty cases of hard classes.



error bound. It encourages larger margins for minority classes and smaller margins for majority classes. Meanwhile, the inputs of LDAM loss should be normalized by normalizing last hidden activation layer and the weight vectors of last fully-connected layer with L2 norm. We only use LDAM loss on our meta-vision model.

Hard Classes Mining. We design a hard class mining (HCM) strategy with the accuracy on train and validation set, then augment them with high-resolution data provided by the host. Specifically, we set the threshold on the train set to 80% and classes whose accuracy under 80% are defined as hard classes. Besides, based on the validation set, we only consider the classes whose samples are more than 50, and the threshold is set to 85%. Based on above two principles, we get 83 hard classes. We manually filter corresponding images and remove some dirty cases like too-small target or low-quality images, as shown in Figure 3. Finally, we complement the rest images with high-resolution ones as provided.

4.1.3. Data Augmentation

Including the traditional data augmentation[26] like random horizontal flip, we also use Mixup [27] and CutMix [28] for models' robustness at a probability of 0.4. It should be pointed out that these two data augmentation methods and LDAM loss are not compatible because of the mixed labels. Besides, we use random erase with a probability of 0.2 and Auto Augment (AA) [29], which searches for improved data augmentation policies automatically, which are over ShearX/Y, TranslateX/Y, Rotate, AutoContrast, etc.

4.2. Open-Set Inference Design

Our model is trained as a closed-set classification task described as above, therefore, inferring methods are innovated and well-designed to tackle the open-set challenge that whether test images belong to the "unknown" class. All data used during inference stage are features and prediction scores of training and test sets extracted by our closed-set training models.

4.2.1. Class-wise Weighted Prototype Classifier

Traditional open-set maximum softmax probability (MSP) method only utilizes test prediction scores to judge unknown class probability, lack of using information in train set. Thus, our paper considers a similarity-based method and proposes our Class-wise Weighted Prototype Classifier (CWPC) by constructing class centers using both features and prediction scores of train set. Specifically, we firstly extract features and prediction scores of all images from train set. Then, we utilize them to compute class centers assigning different weights on samples with the same label instead of taking the features of samples equally. We innovate to apply softmax on the maximum prediction scores of all images with the same label to compute the weight of each sample when computing the class center. The weights for samples of each class can be formulated as follows:

$$\begin{aligned} P_i &= [p_1, p_2, \dots, p_c], \\ m_i &= \text{Max}(P_i), \\ M_c &= [m_1, m_2, \dots, m_{N_c}], \\ W_c &= \text{Softmax}(M_c) \end{aligned} \tag{1}$$

where P_i is the prediction score on i_{th} image after Softmax function, m_i denotes the maximum prediction score of i_{th} image, N_c is the number of images of class c in the training set, M_c denotes prediction scores of N_c images, and W_c denotes weights for images of class c .

CWPC improves the compactness within each class, resulting in more accurate class center representations and achieving powerful prerequisites on subsequent similarity measurements. Also, as an inference-stage strategy, CWPC consumes negligible computation resources, and can be applied on inference stage of all open-set image classification tasks with great generalization, which we consider as a universal algorithm in open-set challenge.

4.2.2. ObservationId-awared Weighted Similarity

To meet the requirements of submission on ObservationId, our paper designs an ObservationId-awared Weighted Similarity (OAWS) to make fusion on images with same ObservationId. As CWPC outputs all class centers, OAWS module aims at calculating the similarity between features of ObservationId and each class center. Thus, we first employ fusion strategy on images with the same ObservationId, where different weights are applied on different images.

The fusion weights for images with the same ObservationId are computed as follows,

$$\begin{aligned}
P_i &= [p_1, p_2, \dots, p_c], \\
m_i &= \text{Max}(P_i), \\
K_o &= [k_1, k_2, \dots, k_{N_o}], \\
W_o &= \text{Softmax}(K_o)
\end{aligned} \tag{2}$$

where P_i is prediction score on i_{th} image after Softmax function, m_i denotes the maximum prediction score of i_{th} image, K_o denotes prediction scores of N_o images with the same ObservationId, W_o denotes weights for images with the same ObservationId.

Then, cosine similarity is subsequently adopted to measure similarity for final results. Specifically, an adjustable threshold is set, the maximum similarity under which belongs to the "unknown" class on test set. OAWS module not only serves as a special technique on Fungi Challenge, but also has referenced significance for open-set challenge for its novelty on similarity and threshold design, which can be further adjusted to achieve better performance.

4.2.3. Comparisons

As CWPC and OAWS successfully generalize closed-set training to open-set recognition, achieving prominent improvements on Fungi Challenge, several comparative methods are proposed. First, based on MSP, we calculate the maximum prediction score and set a threshold to judge whether it belongs to the "unknown" class. It should be noted that as the particularity in Fungi Challenge is ObservationId-format result, we calculate the average test prediction score within each ObservationId as follows,

$$\begin{aligned}
P_i &= [p_1, p_2, \dots, p_c], \\
m_j &= \text{Mean}([P_1, P_2, \dots, P_i]), \\
P_{max} &= \text{Max}(\text{Softmax}(m_j))
\end{aligned} \tag{3}$$

where P_i is prediction score on i_{th} image after Softmax function, m_j represents the mean prediction score of j_{th} ObservationId, P_{max} denotes the maximum prediction score of j_{th} ObservationId.

Second, besides CWPC, class centers can be calculated by using three other selection strategies proposed as follows.

- Average Selection: use average features from all images in the train set to calculate class centers as follows,

$$F_{average_j} = \text{Mean}([f_1, f_2, \dots, f_i]) \tag{4}$$

where f_i is features of i_{th} image in j_{th} classes.

- Filter Selection: use average features from images in the train set whose maximum prediction score is above threshold to calculate class centers.
- GT Selection: use average features from images whose predictions are the same as GT to calculate class centers.

Third, besides OAWS, we apply two other different fusion strategies on test features.

- Average Fusion: test features per ObservationId are the average features of all images with the same ObservationId.
- Filter Fusion: test features per ObservationId are the average features of images whose maximum prediction score is above threshold.

Fourth, we consider using features of every single image in train set instead of class centers. We calculate the similarity between test features of each ObservationId and features of every single image, and extract the top-1 or top-k prediction using model ensemble strategy in Section. 4.2.4, named as Single-Image Similarity Top1 and Single-Image Similarity Top9.

Fifth, we conduct OpenMax, an open-set inference strategy based on Extreme Value Theory (EVT), to estimate the probability of an input being from an unknown class. The key element of estimating the unknown probability is adapting Meta-Recognition concepts to the activation patterns in the penultimate layer of the network.

4.2.4. Inference Augmentation

Test-time Augmentation (TTA). TTA aims at creating multiple enhanced copies of images on test sets, which allows models to make predictions on both original and augmented copies to improve the mean F1 score on the test set. Typical TTA methods such as crop, flip, color jitter are used in Fungi Challenge, where we use random crop with an extension rate of 1.15 on input size, five crop with an additional extension of 32 pixels, horizontal flip with a probability of 1, color jitter with a scope of 0.2, and conduct fusion TTA methods based on above.

Diverse Model Ensemble. As diverse network architectures, training strategies, data augmentations are proposed to improve the performance of models, the variability and diversity between models greatly differs. In order to take full advantage of the semantic information of different models, model ensemble methods are essential to make improvements on final results, which voting is considered as the easiest and most efficient way. We propose top-1 and top-5 voting strategy on diverse models in Fungi Challenge. Top-1 strategy follows "the minority obeys the majority" rule to find the majority class index as the final result. Top-5 strategy extracts predicted top-5 class index of each model and differently weigh them, and chooses the class index with maximum weight as the final result,

$$W_{vote} = [1, 1/2, 1/3, 1/4, 1/5] \quad (5)$$

where W_{vote} is the weight of $Top1$ to $Top5$.

5. Experiments

5.1. Implementation Details

We fine-tune our MetaFormer-2 on ImageNet-21K pre-trained models with input resolution 224×224 on 4 Nvidia T4 GPUs and 384×384 input resolution on 4 Nvidia V100 GPUs. AdamW optimizer is employed with a cosine learning rate scheduler. The learning rate is initialized as 2×10^{-4} for 30 epochs and the first 3 epochs are set for warm-up from 5×10^{-8} . As for vision-only models, we fine-tune SwinTransformer-Base and ConvNeXt networks on 4 Nvidia

Table 1
Results on selection strategies.

Method	Model	Test Input	Macro-F1(%)
MSP	MetaFormer	224×224	75.39
Average Selection	MetaFormer	224×224	75.74
Filter Selection	MetaFormer	224×224	75.58
Average Selection	MetaFormer	384×384	77.39
GT Selection	MetaFormer	384×384	77.24
CWPC	MetaFormer	384×384	77.49

V100 GPUs for 30 epochs. Both are pre-trained on ImageNet-21K and the pretraining weights are provided by the official. The optimizer, learning rate and scheduler is the same as MetaFormer-2 but no warm-up epochs. We choose SwinTransformer-base and ConvNeXt-base with input resolution 384×384 in balance of computational consumption. The weight decay is 10^{-8} for SwinTransformer-base and 2×10^{-5} for others.

5.2. Result

We totally train 7 models, two of which are vision-only models and five are meta-vision models. The evaluation metric for this competition is Mean F1-Score, denoted as Macro-F1, and the results are shown in Tab. 6. We conducted these 7 experiments with different training data, multi-scale input size and loss function. Particularly, test set images with pseudo-labels given by diverse classifiers are used to further fine-tune our trained model. These settings ensure the diversity of models during the model ensemble stage which can complement each other to a better result. Finally, we got 6-th place with a result of 81.02% on public leaderboard and 77.58% on private leaderboard.

5.3. Ablation Studies

We conduct ablation studies to demonstrate the effectiveness of our strategies on selection, fusion, similarity, augmentation and ensemble. Tab. 1 proves that CWPC is the best selection strategy in Fungi Challenge. Tab. 2 proves that OAWS is the best fusion strategy in Fungi Challenge. Tab. 3 and Tab. 5 proves that CWPC and OAWS is the best open-set strategy in Fungi Challenge. Tab. 4 proves that fivecrop is the best test-time augmentation strategy in Fungi Challenge. Tab. 6 proves that Top5 voting is the best ensemble strategy in Fungi Challenge and our ensembled model achieves final result of 81.02% on public leaderboard and 77.58% on private leaderboard.

6. Conclusions

In this paper, we propose a novel open-set fine-grained image classification method called Class-wise Weighted Prototype Classifier (CWPC) using extra text information for FungiCLEF-2022 challenge. We decouple all the process into closed-set training and open-set testing. Thus,

Table 2

Results on fusion strategies.

Method	Model	Test Input	Macro-F1(%)
Average Fusion	MetaFormer	384×384	76.83
Filter Fusion	MetaFormer	384×384	76.36
OAWS	MetaFormer	384×384	77.06

Table 3

Results on similarity strategies.

Method	Model	Test Input	Macro-F1(%)
Single-Image Similarity Top1	SwinTransformer-Base	384×384	64.98
Single-Image Similarity Top9	SwinTransformer-Base	384×384	64.15
CWPC	SwinTransformer-Base	384×384	73.00

Table 4

Results on test-time augmentation strategies.

Test-Time Augmentation	Macro-F1(%)
None	77.49%
Horizontal flip + Vertical Flip + Origin	77.69
Horizontal flip + Color jitter + Origin	77.81
CenterCrop	77.24
RandomCrop	77.71
FiveCrop	78.28

Table 5

Results on overall strategies.

Method	Model	Test Input	Macro-F1(%)
OpenMax	Convnext-Base	384×384	77.18
CWPC + OAWS	Convnext-Base	384×384	79.41

it can benefit from the numerous advances in closed-set image classification, such as large-scale pre-trained models, label smoothing and data augment. It is also cost-saving to generalize closed-set recognition models to open-set scenarios without any further modification with our methods. Besides, we add extra metadata to improve the performance of fine-grained classification using a hybrid structure where convolution is used to extract deep vision features and transformer is used to fuse vision and metadata embedding. With other long-tailed solution and data augmentation, we got 6-th place in this challenge with a final result of 81.02% on public leaderboard and 77.58% on private leaderboard.

Table 6
Final results.

Model	Input size	Train set	Val set	Pseudo	HCM	Loss	Macro-F1(%)
MetaFormer	224×224	✓				CE Loss	78.33
MetaFormer	384×384	✓	✓			LDAM Loss	79.72
MetaFormer	384×384	✓				CE Loss	78.28
MetaFormer	384×384	✓	✓	✓		LDAM Loss	78.22
MetaFormer	384×384	✓	✓		✓	LDAM Loss	79.42
Convnext-Base	384×384	✓	✓	✓		CE Loss	79.81
SwinTransformer-Base	384×384	✓				CE Loss	73.00
Ensemble Top5							81.02

7. Acknowledgments

This work is supported by Chinese National Natural Science Foundation (62076033, U1931202) and MoE-CMCC "Artificial Intelligence" Project No. MCM20190701.

References

- [1] 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.
- [2] 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.
- [3] L. Picek, M. Šulc, J. Heilmann-Clausen, J. Matas, Overview of FungiCLEF 2022: Fungi recognition as an open set classification problem, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.
- [4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, Ieee, 2009, pp. 248–255. doi:10.1109/cvprw.2009.5206848.
- [5] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, S. Belongie, The inaturalist species classification and detection dataset, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8769–8778.
- [6] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
- [7] K. Cao, C. Wei, A. Gaidon, N. Arechiga, T. Ma, Learning imbalanced datasets with label-

- distribution-aware margin loss, in: *Advances in Neural Information Processing Systems*, 2019, pp. 1567–1578.
- [8] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, T. E. Boult, Toward open set recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (2013) 1757–1772. doi:10.1109/TPAMI.2012.256.
 - [9] A. Bendale, T. E. Boult, Towards open set deep networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1563–1572.
 - [10] D.-W. Zhou, H.-J. Ye, D.-C. Zhan, Learning placeholders for open-set recognition, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4401–4410.
 - [11] X. He, Y. Peng, Fine-grained image classification via combining vision and language, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5994–6002.
 - [12] G. Chu, B. Potetz, W. Wang, A. Howard, Y. Song, F. Brucher, T. Leung, H. Adam, Geo-aware networks for fine-grained recognition, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.
 - [13] 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).
 - [14] C. Huang, Y. Li, C. C. Loy, X. Tang, Learning deep representation for imbalanced classification, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5375–5384.
 - [15] Y.-X. Wang, D. Ramanan, M. Hebert, Learning to model the tail, in: *Advances in Neural Information Processing Systems*, 2017, pp. 7029–7039.
 - [16] L. Shen, Z. Lin, Q. Huang, Relay backpropagation for effective learning of deep convolutional neural networks, in: *European conference on computer vision*, Springer, 2016, pp. 467–482.
 - [17] J. Byrd, Z. Lipton, What is the effect of importance weighting in deep learning?, in: *International Conference on Machine Learning*, PMLR, 2019, pp. 872–881.
 - [18] B. Zhou, Q. Cui, X.-S. Wei, Z.-M. Chen, Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9719–9728.
 - [19] X. Wang, L. Lian, Z. Miao, Z. Liu, S. X. Yu, Long-tailed recognition by routing diverse distribution-aware experts, *arXiv preprint arXiv:2010.01809* (2020).
 - [20] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, S. X. Yu, Large-scale long-tailed recognition in an open world, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2537–2546.
 - [21] L. Pícek, 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 Winter Conference on Applications of Computer Vision*, 2022, pp. 1525–1535.
 - [22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
 - [23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).

- [24] 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).
- [25] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.
- [26] S. Marcel, Y. Rodriguez, Torchvision the machine-vision package of torch, in: Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1485–1488.
- [27] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, arXiv preprint arXiv:1710.09412 (2017).
- [28] 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 international conference on computer vision, 2019, pp. 6023–6032.
- [29] C. Lin, M. Guo, C. Li, X. Yuan, W. Wu, J. Yan, D. Lin, W. Ouyang, Online hyper-parameter learning for auto-augmentation strategy, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6579–6588.