OOD-CV Challenge Report

September 18, 2023

1 Team details

- Challenge track: Classification Track
- Team name: USTC-IAT-United
- Team leader name: Jun Yu
- Team leader address, phone number, and email: Department of Automation, University of Science and Technology of China, Hefei, Anhui Province, China; +86-13856070316; harryjun@ustc.edu.cn
- Rest of the team members: Keda Lu, Mohan Jing, Yaohui Zhang
- Team website URL: `https://auto.ustc.edu.cn/2021/0510/c25977a484905/page.htm`
- Affiliation: University of Science and Technology of China
- User names on the OOD-CV Codalab competitions: USTC-IAT-United

- Link to the codes of the solution(s): Please see our attachment.

2 Contribution details

- Title of the contribution : Look beyond the nature of the data: Data-centric approach to solving OOD problems

- General method description: We will describe our approach in three stages. The evaluation results in the first two stages are the results of the tests on the development phase, and in the third stage we perform further optimization for the test set in the final phase.

2.1 Stage 1

2.1.1 Selecting model

First, the organizers restrict the use of pre-trained models trained only with ImageNet-1K, so we exclude a part of pre-trained models based on large-scale datasets such as ImageNet-22K, such as Swin Transformer. In fact, We choose the 3 models(ConvNeXt-L, VOLO-D5 and DeiT-L) of ECCV2022 that we experimented on last year as the base models, and in the training phase, we introduced gradient accumulation for VOLO-D5 and DeiT-L, which can make our models converge more stably and perform better on OOD task.

2.1.2 Data augmentation

We try various automatic data augmentation strategies, as well as some general data augmentation methods, as shown in Table 1. It can be seen that Cutmix+Random Erasing+Color jitter can significantly improve the OOD score, and we decide to use this augmentation combination.

In addition, we also try to simulate the test images in OOD scenes with corruption, we divide corruption into four groups, namely weather,

Table 1: The effect of different augmentation method(development phase).

Augmentation	IID Top-1	OOD Top-1(mean)
RandAugment	+0.14%	-0.19%
Augmix	-0.23%	-0.28%
Random Erasing	-0.18%	+0.34%
Color jitter	+0.32%	+0.37%
Mixup	-0.13%	+0.51%
Cutmix	-0.11%	+1.18%
Cutmix+Random Erasing	-0.16%	+1.38%
Cutmix+Random Erasing+Color jitter	+0.28%	**+1.79**

Table 2: The effect of different corruption method(development phase).

corruption	IID Top-1	OOD Top-1(mean)
Weather	+0.12%	+0.05%
Weather and Digital	+0.35%	**+0.13%**
Weather, Digital and Noise	+0.26%	+0.12%
Weather, Digital, Noise and Blur	+0.21%	+0.10%

digital, noise and blur. Adding Gaussian noise, which is additive noise, and some blur operations hardly improve the generalizability of OOD in realistic scenes. As shown in Table 2, where the combination of Weather + Digital works better.

In the testing stage, we use Test Time Augmentation(TTA) such as FiveCrop.

2.1.3 Adding modules

We try to use Exponential Moving Average (EMA) to mitigate the overfitting, and we add EMA on all three models. as shown in Table 3, the Transformer family of models brings very weak improvement, while the CNN family of models brings a significant improvement.

There are many challenges between the train and test sets of OOD classification dataset, such as unseen distribution and domain shift. Thus we can solve the task which does not have suitable training data to ensure generalization by exploring sample relationships. Among recent

Table 3: The effect of EMA or BF(development phase).

Backbone	OOD Top-1(EMA)	OOD Top-1(BF)
DeiT-L	+0.04%	+0.22%
ConvNeXt-L	+0.49%	+0.18%
VOLO-D5	+0.02%	+0.28%

Table 4: The effect of Single Model(development phase).

Backbone	IID Top-1	OOD Top-1(Mean)
DeiT-L	90.56%	92.81%
ConvNeXt-L	90.84%	92.45%
VOLO-D5	91.09%	93.35%

data scarcity learning methods, sample relationships have been intensively explored using an explicit scheme from either regularization or knowledge transfer. Specifically, a simple yet very effective way is to directly generate new data samples from existing training data, such as mixup, cutmix, copy-paste, crossgrad. Another approach is not to explore sample relationships from the input but to enable the neural network itself to explore sample relationships, such as BatchFormer, which explores sample relationships from a batch perspective. Therefore, we use BatchFormer(BF) to help explore the association between the samples and improve the robustness of the model to identify OOD data. BatchFormer is a model suite that easily loads. We add BatchFormer to overall architecture of the model. As shown in Table 3, the scores of each model show some improvement.

2.1.4 Final Model

We show the best results for each of the 3 models at the development phase, as shown in Table 4.

2.1.5 Post-processing

We perform an exploratory analysis of the confusion matrix obtained from the fused logits of the individual image outputs by image category

calculation. We find several obvious problems that the Chair is easily misclassified as Sofa or Dining table, therefore, we can post-process these two categories from the fused logits. The specific approach can be seen in Algorithm 1, where we take Sofa and Chair as examples and correct the labels according to fused logits.

Algorithm 1 Post-processing

Get all samples with predicted label Sofa as D_{Sofa}
for D_i in D_{Sofa} **do**
  $L \leftarrow D_i$ output on the 3 models with fused logits
  $\alpha \leftarrow$ a parameter > 1
  if $L_{Chari} \times \alpha >= L_{Sofa}$ **then**
    label = Chair
  end if
end for

2.1.6 Detection

We analyze the best results from the previous stage and find that among the 6 categories of data, the scores for shape and context are lower compared to the other 4 categories, so we focus on these 2 categories. For the shape type, we find that sofa and chair have very similar shape and texture, but sofa have the distinctive feature that they have more than two positions, so we use the detected bounding boxes to help us classify them. Without using any additional dataset, we train a Cascade-RCNN based detection model using only the OOD training set and label information. We correct the predictions of the model based on the aspect ratio of the detection results.

In addition, since diningtable and chair often appear together and are also very confusing, and since diningtable tend to be longer compared to chairs, the labels can also be corrected with the help of a detection model.

Table 5: The effect of Model Ensemble(final phase).

Method	IID Top-1	OOD Top-1(Mean)
DeiT-L+ConvNeXt-L+VOLO-D5	90.64%	83.5%
1st round of Pseudo-labeling	91.58%	84.7%
2nd round of Pseudo-labeling	91.94%	85.6%
3rd round of Pseudo-labeling	92.51%	**87.5%**

Table 6: The final results on each OOD index(final phase).

shape	pose	texture	context	weather	occlusion
86.5% (1)	90.32% (2)	77.27%(8)	89.8% (2)	89.8% (1)	97.8% (1)

2.2 Stage 3

2.2.1 Iterative Pseudo-labeling

As shown in Table 5, we first obtain the optimal results by processing using the best solution of the development phase. Then we perform iterative Pseudo-labeling training. We output the prediction confidence of each image for the above best results, and images with confidence $>$ 0.5 are selected and add to the training set to retrain DeiT, ConvNeXt and VOLO; then these 3 models replace the original model to output the new prediction confidence, and images with confidence $>$ 0.8 are selected and add to the training set to retrain DeiT, ConvNeXt and VOLO. Finally, these 3 models replace the original model to output the new prediction confidence, and images with confidence $>$ 0.8 are selected to be added to the training set, and samples obtained by detection and Post-processing corrections are also added to the dataset in order to retrain ConvNeXt and VOLO. And then ensemble ConvNeXt and VOLO to output the final result.

2.2.2 Customized post-processing

We perform post-processing on the more confused categories in pursuit of higher scores. Our final result ranks 1st in Codalab, and the final average OOD score is **88.5%**. The specific indicators are shown in Table 6.

- Description of the particularities of the solutions deployed for each of the tracks : It is worth noting that we used part of the solution for the detection track to effectively make a significant improvement in the classification track.

- References:

1. Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[C] International Conference on Machine Learning. PMLR, 2021: 10347-10357.

2. Yuan L, Hou Q, Jiang Z, et al. Volo: Vision outlooker for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

3. Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C] Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11976-11986.

4. Geirhos R, Rubisch P, Michaelis C, et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness[J]. arXiv preprint arXiv:1811.12231, 2018.

5. Li Y, Yu Q, Tan M, et al. Shape-texture debiased neural network training[J]. arXiv preprint arXiv:2010.05981, 2020.

6. Hou Z, Yu B, Tao D. BatchFormer: Learning to Explore Sample Relationships for Robust Representation Learning[C] Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 7256-7266.

7. Tang Z, Gao Y, Zhu Y, et al. Selfnorm and crossnorm for out-of-distribution robustness[J]. 2020.

8. Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations[J]. arXiv preprint arXiv:1903.12261, 2019.

9. Zhao B, Yu S, Ma W, et al. OOD-CV: A Benchmark for Robustness to Out-of-Distribution Shifts of Individual Nuisances in Natural Images[C] Proceedings of the European Conference on Computer Vision (ECCV), 2022.

10. Zhao B, Wang J, Ma W, et al. OOD-CV-v2: An extended Benchmark for Robustness to Out-of-Distribution Shifts of Individual Nuisances in Natural Images[J]. arXiv preprint arXiv:2304.10266, 2023.

11. Xu L, Ouyang W, Bennamoun M, et al. Multi-class token transformer for weakly supervised semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4310-4319.

- Representative image / diagram of the method(s): As shown in Figure 1, this is the overall framework diagram of our approach.

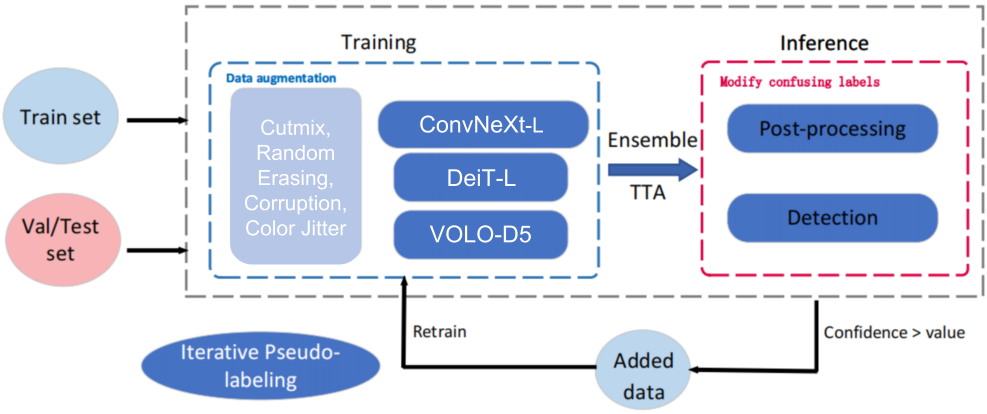

Figure 1: The overall framework diagram of our proposed approach.

3 Global Method Description

- Total method complexity: The project requires the training of 3 classification models as well as a detection model, where the classification model requires 3 iterations and the total complexity should be determined by tripled the VOLO of the classification model with the largest parameters.

- Model Parameters: ConvNeXt-L param count: 196M, DeiT-L param count: 310M, VOLO-D5 param count: 294M.

- Run Time: In the case of 10 V100s, the training takes about 100 hours and the inference takes only half an hour. The training time can be reduced to less than 48 hours when resources allow.

- Which pre-trained or external methods / models have been used: Only the pre-trained model in the ImageNet-1K dataset was used for the experiments.

- Training description : The training description has been quantified and analyzed in detail in Stage 3.

- Testing description: We infer the 3 models obtained by Pseudo-labeling training and perform TTA and post-processing.

- Quantitative and qualitative advantages of the proposed solution : The effect of our approach has been quantified and analyzed in detail in Chapter 2.

- Results of the comparison to other approaches (if any) : The effect of our approach has been quantified and analyzed in detail in Chapter 2.

- Novelty of the solution and if it has been previously published: First we improve the effect of the model. We use models based on different architectures of CNN or Transformer for fusion, which ensures that the model has both global and local inductive bias, which can greatly improve the robustness of the model.

  Secondly, we use BatchFormer to help explore the association between samples and improve the robustness of the model to recognize OOD data. Exploring invariant features between images belonging to the same category also helps in robust representation learning.

In addition, we perform multiple rounds of pseudo-labeling for the fusion of models based on different types of data to ensure that out-of-domain data can be labeled with higher quality, leading to better results.

Finally, our innovations focus on deeper mining of image data, leading to three targeted approaches: using detection to aid classification tasks, style migration based on traditional machine learning methods, and post-processing based on obfuscated category data.

4 Ensembles and fusion strategies

- Describe in detail the use of ensembles and/or fusion strategies (if any).: The fusion method we chose is the fusion of the output layers, where the logits layers of the three models are weighted and fused.

- What was the benefit over the single method? : The model structures we choose are based on CNN or Transformer, respectively. The information of these two types of structures for images is not exactly intersecting, for example, CNN focuses more on local information, while Transformer focuses more on global information, so the fusion can bring a qualitative improvement.

- What were the baseline and the fused methods? : The baseline is a single CNN model, ConvNeXt-L, and the fusion is performed by weighting ConvNeXt-L, DeiT-L, and VOLO-D5 in the ratio of 0.35, 0.3, and 0.35.

5 Technical details

- Language and implementation details (including platform, memory, parallelization requirements) : Project language: Python language Implementation details: four Nvidia V100s with 32G of video memory per gpu. CPU memory is 64G. Convnext is trained in parallel with

two cards, and Deit and VOLO are trained in parallel with four cards.

- Human effort required for implementation, training and validation?: We need to perform deep exploratory data analysis at the beginning of the project implementation, but our approach does not require Human effort for training and validation, and the approach can be deployed end-to-end.

- Training/testing time? Runtime at test per image : Training time: In the case of 10 V100s, it takes up to 100 hours of training, if there are more devices, the fastest training can be completed in 48 hours. Test time: In the case of 10 V100s, it takes only 30 min to infer the final stage of the dataset, with an inference speed of about 20 imgs/s and a time of 0.1s per image tested.

- Comment the efficiency of the proposed solution(s)? : We believe that the solution is still very effective. First, we integrate the most effective CNN and Transformer family of representative models from different architectures and achieve excellent results with only 3 Pseudo-labeling iterations of the model without applying additional datasets.

6 Other details

- General comments and impressions of the OOD-CV challenge. : First of all, we find the OOD-CV challenge very interesting and valuable in solving the current interference with tasks such as recognition and detection in real-life scenarios, and the organizers are very nice and prompt in responding to any questions we asked.

  In addition, we also try to use ECCV 2022 championship solution, MCTformer, to simulate the generation of "Occlusion" data by obtaining foreground images from Imagenet-1K, which is the test set that can significantly improve the "Occlusion" category in the development phase. However, after our testing, the method seems to conflict with our iterative pseudo-labeling in the test phase (since we were able to

get a high score of 0.98 on the "Occlusion" category using only iterative pseudo-labeling), and we are looking into the reasons for this result.

- Other comments: We are willing to explore the nature of OOD data to solve this problem, and we look forward to the announcement of the final results by the organizers. We guarantee that our experimental results are fully reproducible under the pre-trained model using only training data and Imagnet-1k. If the organizers encounter any problems during the reproduction process, please feel free to contact us.