Bag of Tricks for Out-of-Distribution Generalization

Zining Chen¹, Weiqiu Wang¹, Zhicheng Zhao¹,², Aidong Men¹, and Hong Chen³

¹ Beijing University of Posts and Telecommunications
² Beijing Key Laboratory of Network System and Network Culture, China
³ China Mobile Research Institute
{chenzn, wangweiqiu, zhaozc, menad}@bupt.edu.cn
chenhongyj@chinamobile.com

Abstract. Recently, out-of-distribution (OOD) generalization has attracted attention to the robustness and generalization ability of deep learning based models, and accordingly, many strategies have been made to address different aspects related to this issue. However, most existing algorithms for OOD generalization are complicated and specifically designed for certain dataset. To alleviate this problem, nicochallenge-2022 provides NICO++, a large-scale dataset with diverse context information. In this paper, based on systematic analysis of different schemes on NICO++ dataset, we propose a simple but effective learning framework via coupling bag of tricks, including multi-objective framework design, data augmentations, training and inference strategies. Our algorithm is memory-efficient and easily-equipped, without complicated modules and does not require for large pre-trained models. It achieves an excellent performance with Top-1 accuracy of 88.16% on public test set and 75.65% on private test set, and ranks 1st in domain generalization task of nicochallenge-2022.

Keywords: Out-of-Distribution Generalization, Domain Generalization, Image Recognition

1 Introduction

Deep learning based methods usually assume that data in training set and test set are independent and identically distributed (IID). However, in real world scenario, test data may have large distribution shifts to training data, leading to significant decrease on model performance. Thus, how to enable models to tackle data distribution shifts and better recognize out-of-distribution data is a topic of general interest nowadays. Nicochallenge-2022 is a part of ECCV-2022 which aims at facilitating the out-of-distribution generalization in visual recognition, searching for methods to increase model generalization ability, and track 1 mainly focuses on Common Context Generation (Domain Generalization, DG).
Advancements in domain generalization arise from multiple aspects, such as feature learning, data processing and learning strategies. However, as distribution shifts vary between datasets, most of the existing methods have limitations on generalization ability. Especially NICO++ dataset is a large-scale dataset containing 60 classes in track 1, with hard samples including different contexts, multi-object and occlusion problems, etc. Therefore, large distribution shifts between current domain generalization datasets and NICO++ may worsen the effect of existing algorithms.

In this paper, without designs of complicated modules, we systematically explore existing methods which improve the robustness and generalization ability of models. We conduct extensive experiments mainly on four aspects: multi-objective framework design, data augmentations, training and inference strategies. Specifically, we first compare different ways to capture coarse-grained information and adopt coarse-grained semantic labels as one of the objective in our proposed multi-objective framework. Secondly, we explore different data augmentations to increase the diversity of data to avoid overfitting. Then, we design a cyclic multi-scale training strategy, which introduces more variations into the training process to increase model generalization ability. And we find that enlarging input size is also helpful. Moreover, we merge logits of different scales to make multi-scale inference and design weighted Top-5 voting to ensemble different models. Finally, our end-to-end framework with bag of simple and effective tricks, as shown in Figure 2, gives out valuable practical guidelines to improve the robustness and generalization ability of deep learning models. Our solution achieves superior performance on both public and private test set of domain generalization task in nicochallenge-2022, with the result of 88.16% and 75.65% respectively, and ranks 1st in both phases.

2 Related Work

2.1 Domain Generalization

Domain Generalization aims to enable models to generalize well on unknown-distributed target domains by training on source domains. Domain-invariant feature learning develops rapidly in past few years. IRM [1] concentrates on learning an optimal classifier to be identical across different domains, CORAL [17] aims at feature alignment by minimizing second-order statistics of source and target domain distributions. Data processing methods including data generation (e.g. Generative Adversarial Networks [10]) and data augmentation (e.g. Rotation) are simple and useful to increase the diversity of data, which is essential in domain generalization. Other strategies include Fish [15], a multi-task learning strategy that consists the direction of descending gradient between different domains. StableNet [24] aims to extract essential features from different categories and remove irrelevant features and fake associations by using Random Fourier Feature. SWAD [4] figures out that flat minima leads to smaller domain generalization gaps and suffers less from overfitting. Several self-supervised learning methods [12] [7] [3] are also proposed these years to effectively learn intrinsic
image properties and extract domain-invariant representations. Although these methods make great progress on domain generalization, most of them are complicatedly designed and may only benefit on certain dataset.

2.2 Fine-grained Classification

Fine-grained Classification aims to recognize sub-classes under main classes. Difficulty mainly lies in finer granularity for small inter-class variances, large intra-class similarity and different image properties (e.g. angle of view, context and occlusion). Attention mechanisms are mainstream of fine-grained classification which aim at more discriminative foreground features and suppress irrelevant background information [8] [21] [11]. Also, network ensemble methods (e.g. Multiple granularity CNN [20]) including dividing classes into sub-classes or using multi-branch neural networks are proposed. Meanwhile, high-order fine-grained feature is another aspect, which Bilinear CNN [13] uses second-order statistics to fuse context of different channels. However, as fine-grained based methods may have different effects between networks and several with high computational complexity, we only adopt light-weight ECA channel attention mechanism in eca-nfnet-l0 backbone network [2] and SE channel attention mechanism in efficientnet-b4 backbone network [19].

2.3 Generalization Ability

Generalization Ability refers to the adaptability of models on unseen data, which is usually relevant to model overfitting in deep learning based approaches. Reduce model complexity can avoid model fitting into a parameter space only suitable for training set. For example, use models with less parameters and add regularization terms (e.g. L1 and L2 regularization) to limit the complexity of models [9]. Diverse data distribution can also increase generalization ability by using abundant data for pre-training (e.g. Imagenet [6]), applying data augmentation methods [16], and using re-balancing strategies to virtually set differently distributed dataset [27].

3 Challenge Description

3.1 Dataset

The data of domain generalization task in nicochallenge-2022 is from NICO++ dataset [25], a novel domain generalization dataset consisting of 232.4k images for total 80 categories, including 10 common domains and 10 unique domains for each category. The data in domain generalization task is reorganized to 88,866 samples for training, 13,907 for public test and 35,920 for private test with 60 categories. While images from most domains are collected by searching a combination of a category name and a phrase extended from the domain name, there exists hard samples with multi-target, large occlusions and different angle of views, as shown in Figure 1.
Fig. 1. Hard samples with difficult image properties, such as multi-target, occlusions and angle of views, which are easily classified incorrectly for many models.

3.2 Task

Track 1 of nicochallenge-2022 is a common context generation competition on image recognition which aims at facilitating the OOD generalization in visual recognition, whose contexts of training and test data for all categories are aligned and domain labels for training data are available. This task is also known as domain generalization, to perform better generalization ability on unknown test data distribution. Specifically, its difficulty mainly lies in no access to target domains with different distributions during training phase. Thus, the key for this challenge is to improve the robustness and generalization ability of models based on images with diverse context and properties in NICO++ dataset.
4 Method

Our proposed end-to-end framework is illustrated in Figure 2. Firstly, we input multi-scale images based on a cyclic multi-scale training strategy and apply data augmentations to increase the diversity of training data. Then we adopt efficient and light-weight networks (e.g. eca-nfnet-l0 [2]) as backbones to extract features and training with our designed multi-objective head, which can capture coarse-grained and fine-grained information simultaneously. Finally, during inference stage, we merge logits of different scales and design weighted Top-5 voting to ensemble different models.

Fig. 2. Overview of model framework: Input cyclic multi-scale images with data augmentation methods into backbone network to extract features; Extracted Features are fed into multi-objective head to extract coarse-grained and fine-grained logits; Logits are ensembled through TTA methods to output the final result of a single model.

4.1 Multi-objective Framework Design

To capture multi-level semantic information in images, we propose a multi-objective framework. Firstly, domain labels provided by NICO++ dataset naturally contain coarse-grained information and we have considered using them as auxiliary targets to train the backbone network. However, it worsens the performance probably because domain labels focus on the context of images, which may impair the feature learning of foreground objects. Furthermore, we analyse many bad cases, examples from which are illustrated in Figure 3. We find that bicycle is misclassified as horse, wheat is misclassified as monkey and bicycle is
misclassified as gun, respectively, which are far from the correct answer. Therefore, we aim to introduce coarse-grained information to assist model training. Specifically, we manually divide 60 categories into 4 coarse categories according to their properties, denoted as plant, animal, vehicle and object as coarse semantic labels and design a coarse classifier to enable model to learn coarse-grained features. The output dimension is the number of coarse categories, 4, while the output dimension of fine-grained classifier is the number of classes, 60. Under this circumstance, our network can utilize various information from multi-objective, thus increasing robustness and generalization ability of backbone network with barely no computational consumption.

Besides, we also have explored self-supervised objective to increase the generalization ability of models. However, due to large GPU memory consumption and little improvement on test accuracy, we leave detailed self-supervised task design as future work in domain generalization.

4.2 Data Augmentation

Data Augmentation is one of the most significant series of methods in domain generalization, for its simplicity but effectiveness on increasing diversity of data. During training stage, except for common augmentations such as resized-crop, horizontal-flip and normalization, we perform multiple combinations of augmentations and find that the combination of Random Augmentation [5], Cutmix [22] and Mixup [23] and Label Smoothing [18] is the most effective one.

**Random Augmentation** [5] Random Augmentation aims to solve the problem of Auto Augmentation for its high computational cost for the separate search phase on a proxy task, which may be sub-optimal for divergent model and dataset size. It reduces search space for data augmentation and propose one with only 2 hyper-parameters. In this challenge, we set magnitudes, the strength of transformation to 9 and the mean standard deviation to 0.5.

Fig. 3. Examples of bad cases which are easily misclassified as ridiculous results.
Cutmix \[22\] and Mixup \[23\] Cutmix randomly selects two training images and cuts them into patches with the scale of $\sqrt{1 - \gamma}$ on height and width. Then patches from one sample are pasted to another while labels are transformed into one-hot format and mixed proportionally to the area of patches. Mixup also randomly selects two training images and mix them pixel-wise and label-wise with a random number $\lambda$. Both $\gamma$ and $\lambda$ are random numbers, calculated from Beta distribution. In this challenge, we apply these two methods on all batches, with an alternative probability of 0.5. Also, we set $\gamma$ and $\lambda$ to 0.4 and 0.4 by empirical practice, where $\gamma \in [0, 1]$ and $\lambda \in [0, 1]$.

Label Smoothing \[18\] Hard label is prone to overfitting practically in deep learning based approaches, and label smoothing was first proposed to change the ground truth label probability to,

$$p_j = \begin{cases} 
1 - \epsilon, & \text{if } j = y_j, \\
\epsilon/(N - 1), & \text{otherwise}.
\end{cases}$$

(1)

where $\epsilon$ is a constant, $N$ denotes the number of classes, $j$ is the index of class, $y_j$ is the index of ground truth for current image. In this challenge, we set $\epsilon$ to 0.1, where $\epsilon \in [0, 1]$.

Others Except the above methods, we have also exerted Gaussian Blur, Random Erasing and Image-cut, but fail to improve on public test set probably because of conflicts and overlaps between augmentations. For example, Image-cut is a data extension method to cut original images into five images offline, containing four corners and a center one, which has similar effects with multi-scale training and five-crop. Random Erasing may conflict with Cutmix and Mixup for introducing noise on augmented images and Gaussian Blur may impair the quality of images especially with small objects.

4.3 Training Strategy

Different training strategies may lead to severe fluctuations in deep learning based models. In this section, we propose innovative and effective training strategies to enhance the process of model training.

Cyclic Multi-scale Training Due to various scales of objects in NICO++ dataset, we employ Cyclic Multi-scale Training strategy to increase the robustness and generalization ability of our model. Different from multi-scale strategy in object detection which applies multi-scale input in each batch, we propose to change the input size of data periodically for every 5 epochs to better learn representations of objects at different scales, which is suitable for models without pre-training and consume less GPU memory. Also, as we figure out that larger scale is helpful to improve model performance, we set large multi-scales for lightweight eca-nfnet-l0, and small multi-scales for the rest of backbone networks.
**Others** Considering the constraints of GPU memory, we adopt gradient accumulation \cite{14} to increase batch size, which calculates the gradient of a single batch and accumulate for several steps before the update of network parameters and zero-reset of gradient. Besides, we have also verified two-stage training strategies, which CAM \cite{26} is utilized to extract foreground region during second-stage to fine-tune the model. However, little improvement on test set with longer fine-tuning epochs is not worthy.

### 4.4 Inference Strategy

Inference strategies consume little computational resources but may increase model performance significantly with proper design. In this section, we will introduce our multi-scale inference strategy and weighted Top-5 voting ensemble method.

**Test-Time Augmentation** Test-Time Augmentation (TTA) aims to enhance images in test set with proper data augmentation methods and enable models to make predictions on different augmented images to improve model performance. Typical TTA methods including resize, crop, flip, color jitter are used in this challenge, where we first use resize with an extension of 64 pixels on input size. Then we apply different crop strategies, five-crop with an additional extension of 32 pixels, center-crop with an additional extension of 64 pixels. Besides, for center-crop based TTA we use horizontal flip with a probability of 0.5, color jitter with a scope of 0.4, and conduct fused TTA methods based on above. Also, we design multi-scale logits ensemble strategy for multi-scale test. Specifically, we input three different sizes corresponding to different networks, and apply average-weighted (AW) and softmax-weighted (SW), two different ensemble methods to fuse logits of three scales, as shown in Eq. 2 and Eq. 3, respectively. Finally, we compare different TTA combinations to get the best strategy and remove TTAs contradicting with previous strategies (e.g. five-crop and Image-cut).

\[
L_{AW} = [L_1, L_2, L_3] \ast [1/3, 1/3, 1/3]^T
\]

\[
L_{SW} = [L_1, L_2, L_3] \ast \text{Softmax}(\text{Max}(L_1), \text{Max}(L_2), \text{Max}(L_3))^T
\]

where \(L_{AW}\) denotes the ensemble logits by average-weighted method, \(L_{SW}\) denotes the ensemble logits by softmax-weighted method, \(L_i\) denotes the logits of \(i\)-th scale after applying TTA methods.

**Model Ensemble** As diverse model may capture different semantic information due to its unique architecture, model ensemble methods are used to better utilize different context of models to make improvement. Logits ensemble and voting are mainstream methods for their simplicity and efficiency. In this challenge, we propose weighted Top-5 voting strategy on diverse models. Specifically, we get the Top-5 class predictions of each model and then assign voting weights for each
prediction according to its rank. The voting weights for the top-5 predictions of each model can be formulated as Eq. \(4\)

\[
W_{Top5} = [1, 1/2, 1/3, 1/4, 1/5]
\]

(4)

where \(W_{Top5}\) denotes the voting weights from 1-st to 5-th. While voting, we sum the voting weights for the same class prediction from different models and finally take the class prediction with the maximum sum of voting weights as the final result.

5 Experiments

5.1 Implementation Details

Models are trained on 8 Nvidia V100 GPUs, using AdamW optimizer with cosine annealing scheduler. Learning rate is initialized to \(1e^{-3}\) for 300 epochs and weight decay is \(1e^{-3}\) for all models. Batch size is 8 and gradient accumulation is adopted to restrict GPU memory.

5.2 Results

Three backbone networks, eca-nfnet-l0, eca-nfnet-l2 and efficientnet-b4 are trained with cyclic multi-scale training strategy to enrich the diversity of models for better ensemble results. Data augmentations including Cutmix and Mixup, Random Augmentation and Label Smoothing are adopted with empirical hyperparameters to get diverse training data. With inference strategies of TTA and model ensemble, including multi-scale logits ensemble, five-crop and weighted Top-5 voting, we further improve test set performance. During phase 1 on public test set, the evaluation metric is Top-1 accuracy, and finally we rank 1\(^{st}\) with a result of 88.16\%. The results are shown in Table 1.

| Model            | Input size            | Top-1 Accuracy |
|------------------|-----------------------|----------------|
| eca-nfnet-l0     | Multi-scale(large)    | 86.87%         |
| eca-nfnet-l2     | Multi-scale(small)    | 86.55%         |
| efficientnet-b4  | Multi-scale(small)    | 81.43%         |
| ensemble         |                       | 88.16%         |

Table 1. Top-1 accuracy of models on public test set. Multi-scale(small) indicates input size as (448, 384, 320), Multi-scale(large) indicates input size as (768, 640, 512).
5.3 Ablation Studies

We conduct ablation studies to demonstrate the effectiveness of our methods on multi-objective framework design, data augmentations, training and inference strategies, illustrated in Table 2. These methods are added to Baseline step-by-step and effectively improve performance on public test set with negligible computational resources. For example, Cutmix and Mixup improves 12.19%, Multi-scale(small) improves 7.72% and Multi-scale(large) further improves 2.55%. Except the above methods, other strategies basically improve performance for around 2% without any mutual conflicts.

Besides, as mentioned above in Section 4.3, Image-cut has similar effects with multi-scale training and five-crop. Thus, when applying them together, Image-cut can only further improve accuracy of 0.1% with weighted Top-5 voting strategy. CAM based approach can improve accuracy of 0.4% but it requires second-stage training, consuming extra 60 epochs. Therefore, we exclude it from our framework for simplicity. However, the local feature view of Image-cut and the object-sensitive features of CAM based methods are still worth to be explored in future research.

Furthermore, we apply several recent state-of-the-art domain generalization methods, including CORAL, SWAD and StableNet, but they decrease the performance by 0.98%, 2.41%, 1.24% respectively. It further demonstrates that existing algorithms on domain generalization may only benefit on certain dataset and perform worse than heuristic data augmentations.

| Methods                                | Top-1 Accuracy |
|----------------------------------------|----------------|
| Baseline                               | 56.61%         |
| +Multi-scale(small)                    | 64.33%         |
| +Cutmix and Mixup                      | 76.52%         |
| +Random Augmentation                   | 78.92%         |
| +Test-time Augmentation                | 80.53%         |
| +Multi-objective Framework             | 81.53%         |
| +Multi-scale(large) - Multi-scale(small) | 84.08%     |
| +Longer Epochs                         | 86.87%         |
| +Model Ensemble                        | **88.16%**     |

Table 2. Ablation studies on different strategies. Baseline indicates a classic eca-nfnet-l0 backbone network. Except for Model Ensemble, the backbone network of all other strategies is eca-nfnet-l0, and + denotes adding the method based on the previous experimental settings, while − denotes removing the method from the previous experimental settings.
6 Conclusions

In this paper, we comprehensively analyse bag of tricks to tackle image recognition on domain generalization. Methods including multi-objective framework design, data augmentations, training and inference strategies are shown to be effective with negligible extra computational resources. By exerting these methods in a proper way to avoid mutual conflicts, our end-to-end framework consumes low-memory usage, but largely increases robustness and generalization ability, which achieves a significantly high accuracy of 88.16% on public test set and 75.65% on private test set, and ranks 1st in domain generalization task of nicochallenge-2022.

7 Acknowledgments

This work is supported by Chinese National Natural Science Foundation (62076033), and The Key R&D Program of Yunnan Province (202102AE09001902-2).
References

1. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
2. Brock, A., De, S., Smith, S.L., Simonyan, K.: High-performance large-scale image recognition without normalization. In: International Conference on Machine Learning. pp. 1059–1071. PMLR (2021)
3. Carlucci, F.M., D’Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2229–2238 (2019)
4. Cha, J., Chun, S., Lee, K., Cho, H.C., Park, S., Lee, Y., Park, S.: Swad: Domain generalization by seeking flat minima. Advances in Neural Information Processing Systems 34, 22405–22418 (2021)
5. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 702–703 (2020)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Feng, Z., Xu, C., Tao, D.: Self-supervised representation learning from multi-domain data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3245–3255 (2019)
8. Fu, J., Zheng, H., Mei, T.: Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4438–4446 (2017)
9. Goodfellow, I., Bengio, Y., Courville, A.: Regularization for deep learning. Deep learning pp. 216–261 (2016)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems 27 (2014)
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
12. Kim, D., Yoo, Y., Park, S., Kim, J., Lee, J.: Selfreg: Self-supervised contrastive regularization for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9619–9628 (2021)
13. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 1449–1457 (2015)
14. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 (2016)
15. Shi, Y., Seely, J., Torr, P.H., Siddharth, N., Hannun, A., Usunier, N., Synnaeve, G.: Gradient matching for domain generalization. arXiv preprint arXiv:2104.09937 (2021)
16. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of big data 6(1), 1–48 (2019)
17. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: European conference on computer vision. pp. 443–450. Springer (2016)
18. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
19. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)

20. Wang, D., Shen, Z., Shao, J., Zhang, W., Xue, X., Zhang, Z.: Multiple granularity descriptors for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision. pp. 2399–2406 (2015)

21. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2017)

22. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)

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

24. Zhang, X., Cui, P., Xu, R., Zhou, L., He, Y., Shen, Z.: Deep stable learning for out-of-distribution generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5372–5382 (2021)

25. Zhang, X., He, Y., Xu, R., Yu, H., Shen, Z., Cui, P.: Nico++: Towards better benchmarking for domain generalization (2022)

26. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)

27. Zhou, B., Cui, Q., Wei, X.S., Chen, Z.M.: 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. pp. 9719–9728 (2020)