OOD-CV Challenge Report

September 18, 2023

1 Team details

- Challenge track:

  Classification Track - Self-supervised pretrain leaderboard

- Team name:

  Six Walnuts

- Team leader name:

  Kexin Zhang

- Team leader address, phone number, and email:

  Address: No. 2 South Taibai Road, Xian, Shaanxi 710071

  Phone number: +86 18991868563

  Email: ZhangKexin@stu.xidian.edu.cn

- Rest of the team members:

  Rui Peng

  Junpei Zhang

Yuting Yang

Licheng Jiao

Lingling Li

- Team website URL:

  None

- Affiliation:

  School of Artificial Intelligence, Xidian University, Xi'an, China

- User names on the OOD-CV Codalab competitions:

  ZK_

- Link to the codes of the solution(s):

  https://pan.quark.cn/s/a51e565feef2

2 Contribution details

- **Title of the contribution** :

  1st Place Solution for ICCV 2023 Classification Track - Self-supervised pretrain leaderboard

- **General method description**:

  In our solution, we employed EVA and ConvNeXtV2 as baseline models. We enriched the training data with diverse data augmentation methods for six types of out-of-distribution variations, including pose, shape, texture, context, weather, and occlusion. These methods in-

volved mask-level copy-paste, resizing, angle rotation, noise injection, blur, weather simulation, and more.

Initially, we separately utilized two self-supervised pre-trained models, EVA and FCMAE. Subsequently, we fine-tuned the networks with the augmented data. In the next stage, we further trained the models using the improved semi-supervised method, UDA. Throughout the training process, we adhered to the principle of training based on the best convergence to guide the models towards better performance.Finally, we fused the outputs of various models using an optimal consistency-based dynamic fusion approach, resulting in the best submission results.

- **Description of the particularities of the solutions deployed for each of the tracks** :

  Due to the issue of out-of-distribution shifts in the data, we first simulated these six variations (data augmentation) to enrich the training set. In order to enhance the model's generalization and robustness, we incorporated the semi-supervised method UDA into the model for training. We also replaced the cross-entropy loss with label smoothing loss (to enable the model to generalize better to unseen samples and facilitate smoother learning of decision boundaries between classes, thereby improving performance).

- **References**:

  [1] Fang Y, Wang W, Xie B, et al. Eva: Exploring the limits of masked visual representation learning at scale[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

2023: 19358-19369.

[2] 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.

[3] Woo S, Debnath S, Hu R, et al. Convnext v2: Co-designing and scaling convnets with masked autoencoders[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 16133-16142.

[4] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization[J]. arXiv preprint arXiv:1710.09412, 2017.

[5]]Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6022?6031 (2019). https://doi.org/10.1109/ICCV.2019.00612

[6] Xie Q, Dai Z, Hovy E, et al. Unsupervised data augmentation for consistency training[J]. Advances in neural information processing systems, 2020, 33: 6256-6268.

[7]]Zhao, B., Yu, S., Ma, W., Yu, M., Mei, S., Wang, A., He, J., Yuille, A., Kortylewski, A.: Robin: A benchmark for robustness to individual nuisances in real-world outof-distribution shifts. arXiv preprint arXiv:2111.14341 (2021)

[8] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-

Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115, 211?252 (2015)

[9] Chen W, Lin L, Yang S, et al. Self-supervised noisy label learning for source-free unsupervised domain adaptation[C]//2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2022: 10185-10192.

- **Representative image / diagram of the method(s)**:

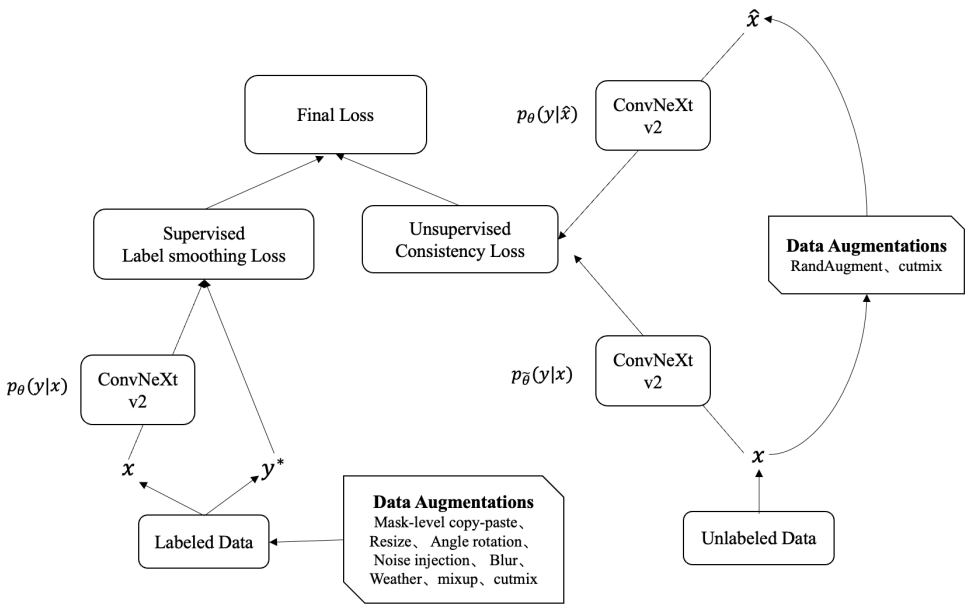

Figure 1: The pipeline of our proposed method.

3 Global Method Description

[* Indicates method used in competition test results.]

- **Total method complexity**:

Model	Flops (G)
EVA-large	236.10
ConvNeXt V2-tiny	76.47
ConvNeXt V2-base	160.38

Table 1: Total method complexity

- **Model Parameters**:

Model	Params (M)
EVA-large	536.53
ConvNeXt V2-tiny	142.64
ConvNeXt V2-base	312.55

Table 2: Model Parameters

- **Run Time**:

Model	Time
EVA-large	56 hours $\times$ 4 GPUs
ConvNeXt V2-tiny	16 hours $\times$ 4 GPUs
ConvNeXt V2-base	38 hours $\times$ 4 GPUs

Table 3: Trainingt Time

- **Which pre-trained or external methods / models have been used:**

Model	Pre-trained model
EVA(Exploring the Limits of Masked Visual Representation Learning at Scale)	EVA
ConvNeXt V2(Co-designing and Scaling ConvNets with Masked Autoencoders)	FCMAE

Table 4: Pre-trained models

- **Training description:**

  Baseline model

  We chose two models as the baseline model :

  1) EVA: A billion parameters vanilla ViT encoder to explore the limits of masked visual representation learning.

  2) ConvNeXt V2: it is specifically designed to be more suitable for self-supervised learning. Using the fully convolutional masked autoencoder pre-training, convnext v2 can significantly improve the performance of pure ConvNets across various downstream tasks.

  Training Process

  1) We applied extensive data augmentation to the model, enlarging the original training dataset by a factor of 10. Techniques such as mask-level copy-paste, resizing, angle rotation, introducing noise, blurring, and simulating various weather conditions were employed to emulate the distribution shift seen in the test set. Additionally, we also utilized two common data augmentation strategies, mixup and CutMix.

  2) We employed the powerful pre-trained models, EVA and Convnext

v2, separately. When training with the augmented training data, due to the unique data distribution, we opted to fine-tune all layers. Although this extended the training time, it yielded superior results. Moreover, given the instability and randomness inherent in the training process, we imposed constraints on the model training following the principle of achieving the best convergence. The approach in the code was as follows: if the model's performance on the validation set did not improve or even declined continuously for three consecutive epochs, the training would be restarted from the epoch with the highest previously saved validation score.

3) Due to the limited amount of training data and the distribution shift, we incorporated the semi-supervised UDA (Unsupervised Data Augmentation for Consistency Training) method into the model's training. The idea of UDA is to utilize both unlabeled and labeled data, employing data augmentation and consistency loss during model training to enhance the model's generalization performance. We also replaced the cross-entropy loss with label smooth loss, a method that proved to be highly effective in enhancing the performance on the test set.

Implementation details: ConvNeXt v2-base is trained with 4 GPUs and 16 samples per GPU. (EVA-large is trained with 4 GPUs and 16 samples per GPU. ConvNeXt v2-tiny is trained with 4 GPUs and 32 samples per GPU.) We use the AdamW optimizer . The learning rate is adjusted according to the cosine decaying policy and the initial learning rate is set to1e-3. The warm-up strategy is applied over the first 20 epochs, gradually increasing the learning rate linearly from 1e-5 to the

initial value of the cosine schedule. The model is trained for 300 epochs. We set β=0.6 and τ=0.4 for UDA.

- **Testing description**:

  Implementation details: ConvNeXt v2-base(EVA-large, ConvNeXt v2-tiny) were tested with 2 GPUs and 32 samples per GPU.

- **Quantitative and qualitative advantages of the proposed solution** :

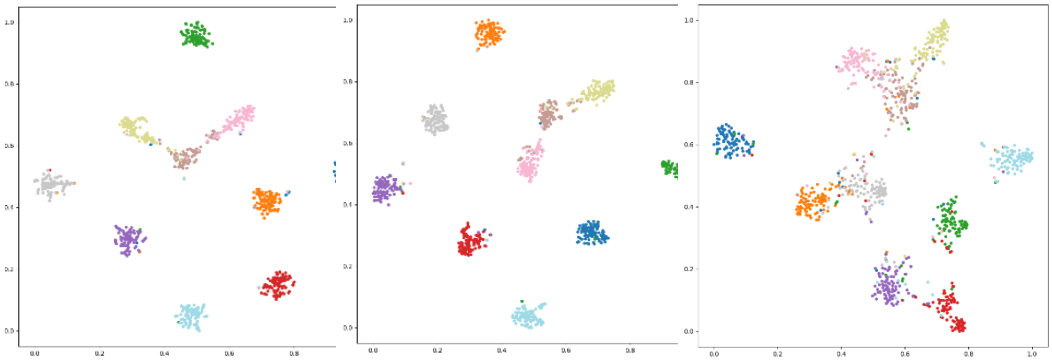

Figure 2: (a) t-SNE (EVA-large) (b) t-SNE (ConvNeXt V2-tiny) (c) t-SNE (ConvNeXt V2-base) Different models' t-SNE visualization results on the validation set. The ConvNeXt V2 model demonstrates a better ability to distinguish between these ten classes. Additionally, ConvNeXt V2 is more effectively trained, yielding better results with the same resources. This suggests that after incorporating semi-supervised training, the ConvNeXt V2 model can adapt more effectively to data shifts.

Methods	shape	pose	context	texture	occlusion	weather
EVA(large)	85.33	89.46	83.52	91.21	82.37	88.25
EVA(large)+DA	90.72	92.25	90.56	94.66	91.28	91.20
EVA(large)+DA +UDA	94.32	97.05	97.78	98.86	97.28	96.04
ConvNeXt V2(tiny) +DA	89.56	91.32	90.23	94.45	91.36	90.89
ConvNeXt V2(tiny) +DA+UDA	93.83	96.93	97.16	97.73	97.56	96.62
ConvNeXt V2(base) +DA	91.08	92.86	91.32	95.16	92.42	91.44
ConvNeXt V2(base) +DA+UDA	95.58	98.24	98.72	98.67	98.63	96.72

Table 5: Ablation Experiments on the Validation Set, where "DA" represents data augmentation, and "UDA" represents semi-supervised training. After introducing data augmentation, the model's performance has improved across various offset categories, especially in terms of content and occlusion. Additionally, the use of semi-supervised training methods has enabled the model to better adapt to data distribution shifts.

- **Results of the comparison to other approaches (if any)** :

  None

- **Novelty of the solution and if it has been previously published**:

  We incorporate a semi-supervised approach (UDA) into the model

training process, further alleviating data distribution shifts and enhancing the model's robustness.

4 Ensembles and fusion strategies

- **Describe in detail the use of ensembles and/or fusion strategies (if any).**:

  We designed an optimal consistency dynamic fusion method to select the best model for fusion. For the test set images, a portion of them is chosen for data augmentation in batches. From these batches, the models that exhibit the best consistency (where predictions remain consistent for samples generated from the same data augmentation) are selected for fusion.

- **What was the benefit over the single method?** :

  We fused the best model among three baseline models, which can enhance the performance on the test set. By combining information from different models, the more models we fuse, the better the results become.

- **What were the baseline and the fused methods?** :

  We fused one model from EVA-large and two models from ConvNeXt v2 tiny, along with two models from ConvNeXt v2 base. After ensemble, the OOD accuracy improved from 91.24% to 91.76%.

5 Technical details

- **Language and implementation details (including platform, memory, parallelization requirements)** :

  Python , PyTorch , 4 × V100 GPUs.

  Our codes built on the MMPretrain platform, easy integration and transferability with other tasks within the OpenMMLab ecosystem, not limited to image classification tasks.

- **Human effort required for implementation, training and validation?**:

  None

- **Training/testing time? Runtime at test per image** :

Methods	Training Time	Testing Time	Runtime at test per image
EVA(large)	56hours × 4GPUs	32min	0.113s
ConvNeXt V2(tiny)	16hours × 4GPUs	12min	0.042s
ConvNeXt V2(base)	38hours × 4GPUs	20min	0.071s

Table 6: Training, testing time and runtime at test per image

- **Comment the efficiency of the proposed solution(s)?** : The primary model is ConvNeXt v2, which is relatively resource-friendly compared to other models.

6 Other details

- General comments and impressions of the OOD-CV challenge. : Thanks to the organizers of the OOD-CV challenge
- Other comments: None.