OOD-CV Challenge Report

September 9, 2024

1 Team details

- Challenge track: OOD-CV Workshop SSB Challenge 2024 (Open-Set Recognition Track)
- Team name: Intellindust-AI-Lab
- Team leader name: Yang Li
- Team leader address, phone number, and email:

  Address: Room 602, Block B, Building 4, Software industry Base, Yuhai Street, Nanshan District, Shenzhen, China;

  Phone number: 0086 15855951043;

  Email: liyang@intellindust.com
- Rest of the team members: Youyang Sha, Shengliang Wu, Yuting Li, Xuanlong Yu, Shihua Huang, Xiaodong Cun, Yingyi Chen, Dexiong Chen, Xi Shen
- Team website URL: https://intellindust.cn/
- Affiliation: Intellindust, Great Bay University, Institute for Research in Biomedicine Bellinzona, Max Planck Institute of Biochemistry
- User names on the OOD-CV Codalab competitions: Intellindust-AI-Lab
- Link to the codes of the solution(s): https://github.com/LIYangggggg/SSB-OSR

2 Contribution details

- Title of the contribution: SURE-OOD
- General method description: SURE-OOD builds on our recent work SURE [1], which was initially developed to train reliable and robust classifiers. SURE integrates various techniques to enhance the reliability and robustness of neural networks. By combining SURE with typical post-hoc methods such as GradNorm [2] and RePlacing (RP) uniform distribution to class-prior distribution [3], we develop SURE-OOD, a highly effective approach to OOD detection. We incorporate standard engineering tricks to further improve performance: Test-Time Augmentation (TTA) and model ensembling.
- Description of the particularities of the solutions deployed for each of the tracks: We employed the DeiT-3 model [4] as our backbone, specifically utilizing the Base variant (DeiT-B, resolution 384) pre-trained on ImageNet-1K [5][1]. Due to computational constraints, we focused on fine-tuning the model using the ImageNet-1K training set [5] to enhance its performance for out-of-distribution (OOD) detection. Our approach integrates several advanced techniques: Sharpness-Aware Minimization (SAM) [6] to identify flatter minima, Stochastic Weight Averaging (SWA) [7] to stabilize model updates, RegMixup [8] for regularized data augmentation, and Cosine Similarity Classifier (CSC) [9] to learn compact feature representations. These methods, which have been demonstrated to be effective in SURE [1], contributed to building a more reliable and robust neural network.

  Additionally, we applied GradNorm [2] but RePlacing the standard uniform distribution by the training set's prior distribution [3] (RP) to achieve more accurate OOD sample quantification. Notably, we found that removing Layer Normalization [10] improved OOD detection performance. Furthermore, we incorporated standard engineering practices such as Test-Time Augmentation (TTA) using five-crop augmentation and model ensembling to optimize performance further.
- References: The references are included at the end of this document.

[1] https://dl.fbaipublicfiles.com/deit/deit_3_base_384_1k.pth

- Representative image/diagram of the method(s): Figure 1 provides a comprehensive depiction of the training and testing processes, illustrating the overall framework and flowchart of the algorithm.

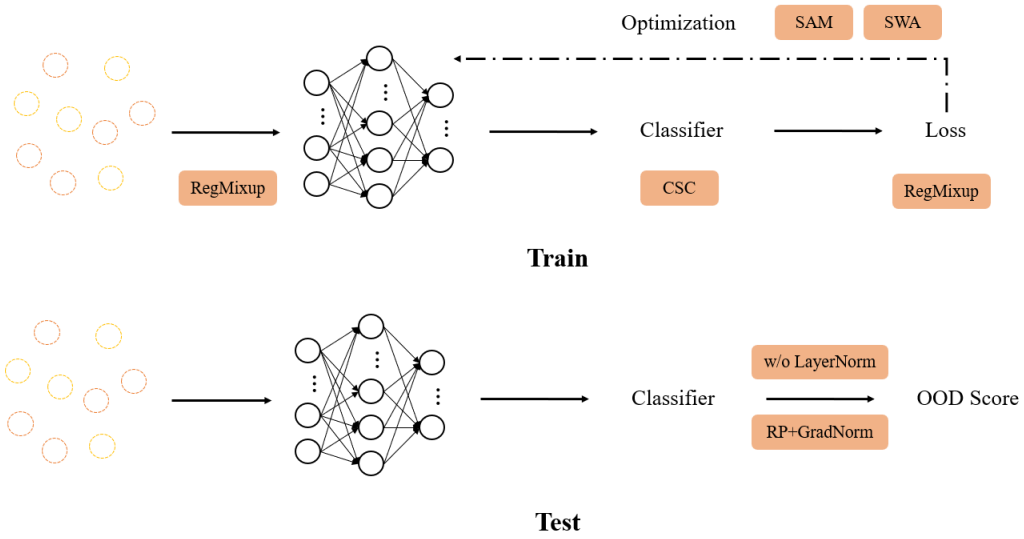

Figure 1: **Training and testing workflow of SURE-OOD.** We utilize the SURE framework for training a classifier. During testing, the OOD score is computed using GradNorm [2], RePlacing the standard uniform distribution by the training set's prior distribution [3] (RP) to achieve more accurate OOD sample quantification. Additionally, we found that removing the final Layer Normalization [10] plays a crucial role in improving OOD detection performance.

3 Global Method Description

[* Indicates the method used in competition test results.]

- Total method complexity: The "deit3-base-patch16-384.fb-in1k" model with a batch size of 128, requires 19,803 MB of peak memory and operates at 92.10 GFLOPs.
- Model Parameters: "deit3-base-patch16-384.fb-in1k" with 87.1MB model parameter.

- Run Time: When testing the "deit3-base-patch16-384.fb-in1k" model on an NVIDIA GeForce RTX 3090 GPU (24GB), it took 21 minutes and 35 seconds to complete one round of testing. When including Test-Time Augmentation (TTA), the testing time increased to approximately 1 hour and 45 minutes.

- Which pre-trained or external methods/models have been used: We used the DeiT-B [4] model, pre-trained on the ImageNet-1K [5] training set, which is available from its official webpage[2].

- Training Description: Following DeiT-3 [4], we employed a range of data augmentation techniques, including resizing, cropping, flipping, color jittering, Gaussian blurring, and solarization. Furthermore, we incorporated Mixup as a regularization strategy, consistent with the methodology outlined in SURE [1]. For learning rate scheduling, we adopted a staged approach, utilizing a cosine learning rate schedule [11] in conjunction with Stochastic Weight Averaging (SWA) [7] to improve training stability.

- Testing description: During inference, we utilized Test-Time Augmentation (TTA), where the input image was processed using five crops (the four corners and the center). Model parameters were averaged from two independent training runs, both conducted with identical hyper-parameters. The image size during training was set to 384×384, while for testing, we increased the input resolution to 480×480. Additionally, the removal of the Layer Normalization [10] before the classification head during testing resulted in a significant improvement in performance.

- Quantitative and qualitative advantages of the proposed solution: In Figure 2, we present the distribution of OOD scores for both in-distribution (ID) and out-of-distribution (OOD) samples. We compare the official DeiT-B model[3] with the same model fine-tuned using SURE [1]. The results clearly demonstrate that fine-tuning with SURE leads to a significant improvement in OOD detection performance.

[2] https://github.com/facebookresearch/deit

[3] https://dl.fbaipublicfiles.com/deit/deit_3_base_384_1k.pth

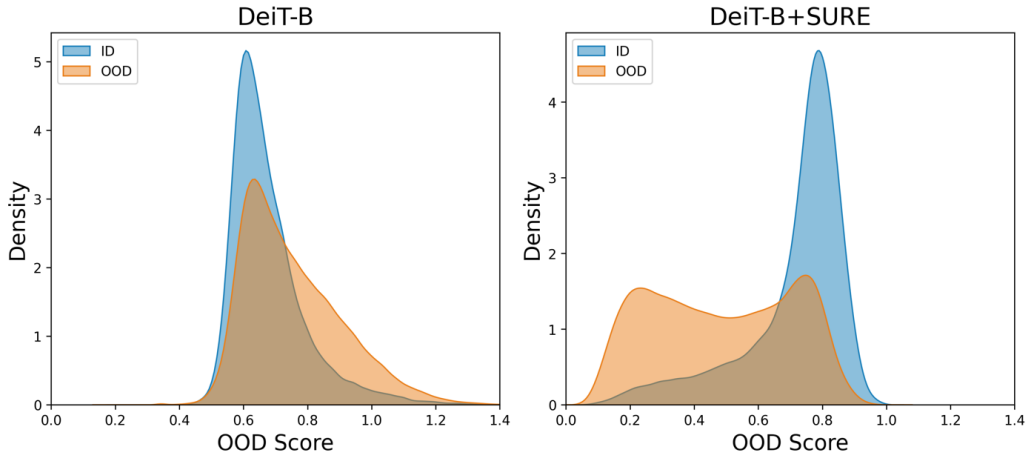

Figure 2: **The distribution of OOD scores for both in-distribution (ID) and out-of-distribution (OOD) samples.** Left: the official DeiT-B model; Right: DeiT-B model finetuning with SURE [1].

- Results of the comparison to other approaches (if any): We present our results across three tables for comprehensive analysis. Table 1 compares the OOD detection performance between the baseline official DeiT-B model, and the SURE fine-tuned model. Table 2 evaluates the effectiveness of various post-hoc methods applied within the SURE framework. Finally, Table 3 examines the performance improvements resulting from post-processing, providing a detailed breakdown of each post-processing step and quantifying its contribution to the overall enhancement of the model's performance.

Table 1: **OOD detection performance on the test dataset provided by SSB challenge for the official DeiT-B model and the SURE fine-tuned one.** All models employ RP+GradNorm [3] for OOD scoring.

Methods	AUROC↑	FPR@TPR95↓
DeiT-B [4]	34.28	96.86
DeiT-B [4]+SURE [1]	**79.97**	**68.03**

- Novelty of the solution and if it has been previously published: We found that integrating the SURE framework with the

Table 2: **OOD detection performance on the test dataset provided by SSB challenge for different post-hoc approaches.** All models are based on the SURE [1] fine-tuned DeiT-B [4]. In this comparison, RP [3] represents the RePlacing strategy, where the uniform distribution used in GradNorm is replaced with the ID class-prior distribution. On the other hand, RW [3] denotes the ReWeighting approach, which reweights the scores according to the similarity between the ID class-prior distribution and the softmax output.

Methods	AUROC↑	FPR@TPR95↓
MSP [12]	76.09	73.02
ODIN [13]	77.79	71.76
Energy [14]	77.16	70.01
GradNorm [2]	57.29	70.35
RP + MSP [3]	76.12	73.19
RW + ODIN [3]	76.49	71.88
RW + Energy [3]	77.71	71.08
RP + GradNorm [3]	**79.97**	**68.03**

"RP+GradNorm" [3] metric yields highly effective OOD detection, a novel approach not previously explored in the literature. Additionally, our results indicate that removing Layer Normalization [10] significantly enhances OOD detection performance.

4 Ensembles and fusion strategies

- Describe in detail the use of ensembles and/or fusion strategies (if any): We employed a model ensembling technique where the parameters of two trained models were averaged.
- What was the benefit over the single method?: As shown in Table 3, AUROC increased by 0.3 points, and the FPR@TPR95 decreased by approximately 0.6 points.
- What were the baseline and the fused methods? We fused two DeiT-B models, each trained independently using the SURE framework, to enhance overall performance.

Table 3: **OOD detection performance on the SSB challenge test dataset using various post-processing techniques.** All models are based on the SURE [1] fine-tuned DeiT-B [4], utilizing "RP+GradNorm"[3] for OOD scoring. LN, Ensemble, and FCrop represent Layer Normalization[10], Model Ensembling, and Five Crop augmentation, respectively.

Methods	AUROC↑	FPR@TPR95↓
SURE	79.97	68.03
SURE w/o LN	80.25	64.59
SURE(480p) w/o LN	80.31	64.25
SURE(480p) w/o LN + Ensemble	80.61	63.58
SURE(480p) w/o LN + Ensemble + FCrop	**81.54**	**61.72**

5 Technical details

- Language and implementation details (including platform, memory, parallelization requirements): We implemented our method in PyTorch. Training was conducted on four NVIDIA GeForce RTX 3090 GPUs (24GB each), while testing was performed on a single GPU.

- Human effort required for implementation, training and validation?: The primary human efforts involved include downloading and preprocessing data, modifying training code to suit specific needs, and experimenting with various post-processing methods to optimize the model's performance.

- Training/testing time? Runtime at test per image: For the model "deit3-base-patch16-384.fb-in1k", the runtime at test is approximately 176 images per second. The total training time is about 20 hours, and the testing phase takes around two and a half hours in total.

- Comment the efficiency of the proposed solution(s)?: Despite limited training resources (four NVIDIA GeForce RTX 3090 GPUs with 24GB memory), our team achieved strong performance. We attribute a portion of this success to utilizing the pre-trained model, which played a significant role in enhancing the results. We believe that the proposed solution has significant potential for further improvement, particularly with the emergence of more advanced models.

6 Other details

- General comments and impressions of the OOD-CV challenge: The platform provided by the OOD-CV challenge has been highly appreciated, offering a valuable opportunity to showcase and evaluate our methods. It is hoped that the challenge will continue to grow and improve in the future.

- Other comments: Special thanks are extended to Bingchen Zhao for his patience and support, as his detailed responses to the issues encountered during the competition were invaluable.

References

[1] Yuting Li et al. "SURE: SUrvey REcipes for building reliable and robust deep networks". In: *CVPR*. 2024 (cit. on pp. 2, 4–7).

[2] Rui Huang, Andrew Geng, and Yixuan Li. "On the Importance of Gradients for Detecting Distributional Shifts in the Wild". In: *NeurIPS*. 2021 (cit. on pp. 2, 3, 6).

[3] Xue Jiang et al. "Detecting out-of-distribution data through in-distribution class prior". In: *ICML*. 2023 (cit. on pp. 2, 3, 5–7).

[4] Hugo Touvron, Matthieu Cord, and Hervé Jégou. "DeiT III: Revenge of the ViT". In: *ECCV*. 2022 (cit. on pp. 2, 4–7).

[5] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *CVPR*. 2009 (cit. on pp. 2, 4).

[6] Pierre Foret et al. "Sharpness-aware minimization for efficiently improving generalization". In: *ICLR*. 2020 (cit. on p. 2).

[7] Pavel Izmailov et al. "Averaging weights leads to wider optima and better generalization". In: *Conference on Uncertainty in Artificial Intelligence (UAI)*. 2018 (cit. on pp. 2, 4).

[8] Hongyi Zhang et al. "mixup: Beyond empirical risk minimization". In: *ICLR*. 2018 (cit. on p. 2).

[9] Francesco Pinto et al. "Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness". In: *NeurIPS*. 2022 (cit. on p. 2).

[10] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. "Layer normalization". In: *ArXiv* (2016) (cit. on pp. 2–4, 6, 7).

[11] Ilya Loshchilov and Frank Hutter. "Sgdr: Stochastic gradient descent with warm restarts". In: *ICLR*. 2017 (cit. on p. 4).

[12] Dan Hendrycks and Kevin Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks". In: *ICLR*. 2017 (cit. on p. 6).

[13] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. "Enhancing the reliability of out-of-distribution image detection in neural networks". In: *ICLR*. 2018 (cit. on p. 6).

[14] Weitang Liu et al. “Energy-based out-of-distribution detection”. In: *NeurIPS*. 2020 (cit. on p. 6).