Supplementary Material:
High-Rate Mixout: Revisiting Mixout for Robust Domain Generalization

Masih Aminbeidokhti Heitor Rapela Medeiros Srikanth Muralidharan
Eric Granger Marco Pedersoli
École de technologie supérieure, Montreal, Canada

A. Computational and Memory Analysis

We analyze memory usage and FLOPs for each method, where total FLOPs per iteration include forward and backward passes. The backward pass computes weight gradients (∇W) and input gradients (∇X), each requiring $\sim 1\times$ the FLOPs of the forward pass.

A.1. Single-Run Methods

For conventional methods (e.g., ERM):

$$\text{FLOPs} = 3 \times \text{Forward FLOPs} = \begin{cases} 12.3\,\text{G} & (\text{ResNet50: } 4.1 \times 3), \\ 13.8\,\text{G} & (\text{ViT-S/16: } 4.6 \times 3). \end{cases}$$

We need to store 100% of weight and activation gradients.

A.2. Ensemble-Based Methods

For 18-model ensembles:

$$\text{FLOPs} = 18 \times \text{ERM FLOPs}, \quad \text{Memory} = \text{ERM Memory (sequential training)}.$$

A.3. Mixout

For swap rate $p\%$:

- **FLOPs**: Skips $p\%$ of ∇W computation:

$$\text{FLOPs} = \text{Base} \times \left(2 + \frac{p}{100}\right), \quad \text{where Base} = \text{Forward FLOPs}.$$

- **Memory**: Reduces storage by $p\%$:

$$\text{Memory} = \left(1 - \frac{p}{100}\right) \times \text{Full Memory}.$$

  Input gradient computations remain unchanged.

A.4. LoRA for ViT-S/16

Forward Pass: Replacing all linear layers in ViT-S/16 with LoRA of rank 64 adds the following FLOPs:

$$\text{FLOPs}_{\text{LoRA-forward}} = 12 \cdot N \cdot \left[4 \cdot (D \cdot r + r \cdot D) + 2 \cdot (D \cdot r + r \cdot 4D)\right],$$

where $N = 196$ (sequence length), $D = 384$ (hidden dim), $4D = 1536$ (MLP intermediate dim), and $r = 64$. Substituting values:

$$\text{FLOPs}_{\text{LoRA-forward}} = 12 \cdot 196 \cdot [4 \cdot (384 \cdot 64 + 64 \cdot 384) + 2 \cdot (384 \cdot 64 + 64 \cdot 1536)] = 1.04\,\text{GFLOPs}.$$

Total forward FLOPs:

$$4.6\,\text{GFLOPs} + 1.04\,\text{GFLOPs} = \mathbf{5.64}\,\text{GFLOPs}.$$

Backward Pass: LoRA backward FLOPs are twice the forward FLOPs (gradient computations for A and B):

$$\text{FLOPs}_{\text{LoRA-backward}} = 2 \cdot \text{FLOPs}_{\text{LoRA-forward}} = 2.08\,\text{GFLOPs}.$$

Final FLOPs is

$$9.2\,\text{GFLOPs} - 4.6\,\text{GFLOPs} + 2.08\,\text{GFLOPs} = \mathbf{6.68}\,\text{GFLOPs}.$$

B. Implementation details

Following DomainBed benchmark [4], we evaluate our method on five diverse datasets. PACS [6], VLCS [3], Office Home [8], TerraIncognita [1] and DomainNet [7]. We report out-of-domain accuracies for each domain and their average using a leave-one-out cross-validation method. We use Adam [5] optimizer with a mini-batch containing all domains and 32 examples per domain. We follow [2] and train models for 15000 steps on DomainNet and 5000 steps for other datasets, corresponding to a variable number of epochs dependent on dataset size. Every experiment is repeated three times with different seeds. We leave 20% of the source domain data for validation. For each domain within a dataset, we follow the procedure described in [2], sampling 18 hyperparameter sets from the search space and fine-tuning ERM to identify the best configuration. We then conduct a separate hyperparameter search tailored to each method. Each domain is sequentially used as the target (test) domain, while the remaining domains are utilized as source (training) domains (please refer to the appendix for more details on datasets and hyperparameter configurations).

Table 1. Hyperparameters used for all methods in and their respective distributions for grid search.

Hyperparameter	**Search Space**
batch size	32
learning rate	{1e-5, 3e-5, 5e-5}
classifier dropout	{0.0, 0.1, 0.5}
weight decay	{1e-4, 1e-6}
swap rate	{0.1, 0.2, $\cdots$, 0.9 }

B.1. Datasets

PACS: is a 7-way object classification task with 4 domains: art, cartoon, photo, and sketch, with 9,991 samples [6].

VLCS: is a 5-way classification task from 4 domains: Caltech101, LabelMe, SUN09, and VOC2007. There are 10,729 samples. This dataset mostly contains real photos. The distribution shifts are subtle and simulate real-life scenarios well [3].

Office Home: is a 65-way classification task depicting everyday objects from 4 domains: art, clipart, product, and real, with a total of 15,588 samples [8].

TerraIncognita: is a 10-way classification problem of animals in wildlife cameras, where the 4 domains are different locations, L100, L38, L43, L46. There are 24,788 samples. This represents a realistic use case where generalization is indeed critical [1].

DomainNet: is a 345-way object classification task from 6 domains: clipart, infograph, painting, quickdraw, real, and sketch. With a total of 586,575 samples, it is larger than most of the other evaluated datasets in both samples and classes [7].

C. Full Results

In this section, we show detailed results of Table 1 of the main manuscript. Tables 2, 3, 4, 5 6 show full results on PACS, VLCS, OfficeHome, TerraIncognita, and DomainNet datasets, respectively. The provided tables summarize the obtained out-of-distribution accuracy for every domain within the five datasets. Standard deviations are reported with different seeds when possible. To guarantee the comparability of the results, we followed the same experimental setting as in DomainBed [4].

References

[1] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV*, pages 456–473, 2018. 2

[2] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Neurips*, 34:22405–22418, 2021. 2

[3] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *ICCV*, pages 1657–1664, 2013. 2

[4] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 2, 3

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[6] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5542–5550, 2017. 2

[7] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415, 2019. 2

[8] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017. 2

Table 2. OOD accuracies on PACS.

Method	A	C	P	S	**Avg.**
ResNet50					
Multi-run training (18 Models)					
ENS	90.85	83.53	98.88	82.95	89.05
DiWA	92.01	84.01	99.18	81.65	89.21
Single-run training					
CORAL	89.36±0.76	80.44±0.99	98.58±0.11	81.23±0.82	87.40±0.67
Large Dropout	87.78±1.31	82.68±0.28	98.43±0.15	79.66±0.75	87.14±0.62
ERM	90.81±0.87	81.68±0.78	98.68±0.26	79.45±0.94	87.66±0.71
High-rate Mixout	91.07±0.14	83.46±0.18	99.15±0.04	79.32±0.50	88.25±0.22
ViT-S/16					
Multi-run training (18 Models)					
ENS	90.30	82.20	99.10	79.96	87.89
DiWA	90.79	83.00	99.25	80.47	88.38
Single-run training					
ERM	87.27±0.60	81.11±0.96	98.30±0.11	77.47±1.03	86.04±0.68
High-rate Mixout	89.91±0.47	80.99±0.37	98.83±0.05	77.42±1.21	86.79±0.52

Table 3. OOD accuracies on VLCS.

Method	C	L	S	V	**Avg.**
ResNet50					
Multi-run training (18 Models)					
ENS	98.06	64.89	76.28	80.90	80.03
DiWA	98.06	63.67	76.96	89.64	79.83
Single-run training					
CORAL	98.82±0.10	64.94±0.69	76.83±0.77	79.46±0.52	80.01±0.52
Large Dropout	97.76±0.46	64.82±0.35	74.41±0.38	80.25±0.54	79.31±0.43
ERM	98.06±0.15	64.28±0.49	76.72±0.48	79.48±0.60	79.64±0.43
High-rate Mixout	98.20±0.10	65.68±0.12	73.88±0.55	79.85±0.80	79.40±0.39
ViT-S/16					
Multi-run training (18 Models)					
ENS	97.26	65.65	77.53	83.38	80.96
DiWA	96.64	64.85	77.68	82.90	80.52
Single-run training					
ERM	96.91±0.14	64.49±0.22	75.65±0.93	82.29±0.13	79.83±0.36
High-rate Mixout	95.88±0.46	64.74±0.46	76.96±0.20	79.49±0.57	79.27±0.42

Table 4. OOD accuracies on OfficeHome.

Method	A	C	P	R	**Avg.**
ResNet50					
Multi-run training (18 Models)					
DiWA	70.55	53.64	79.76	83.02	71.74
ENS	69.77	54.04	79.95	83.33	71.77
Single-run training					
CORAL	70.08±0.50	53.20±0.39	78.95±0.29	82.69±0.18	71.23±0.34
Large Dropout	68.64±0.74	53.30±0.30	78.13±0.43	82.56±0.05	70.66±0.38
ERM	68.95±1.16	52.13±0.67	78.61±0.48	82.14±0.49	70.46±0.70
High-rate Mixout	71.06±0.68	54.20±0.35	79.99±0.11	83.30±0.06	72.14±0.30
ViT-S/16					
Multi-run training (18 Models)					
ENS	71.83	56.10	81.42	82.62	72.99
DiWA	72.40	55.87	81.17	82.36	72.95
Single-run training					
ERM	68.74±0.44	54.16±0.61	80.01±0.27	81.64±0.20	71.14±0.38
High-rate Mixout	72.11±0.10	55.05±0.32	80.78±0.11	82.36±0.18	72.58±0.18

Table 5. OOD accuracies on TerraIncognita.

Method	L100	L38	L43	L46	**Avg.**
ResNet50					
Multi-run training (18 Models)					
ENS	63.67	46.44	63.48	42.83	54.10
DiWA	62.98	50.44	62.47	46.85	55.68
Single-run training					
CORAL	58.21±1.94	47.59±1.62	57.65±0.51	38.98±1.84	50.61±1.48
ERM	59.53±2.79	48.93±1.79	61.87±1.57	40.13±3.17	52.62±2.33
Large Dropout	61.31±2.79	47.65±1.99	60.71±0.64	39.41±0.77	52.27±1.55
High-rate Mixout	65.53±0.49	56.93±1.32	64.27±0.21	46.95±0.64	58.42±0.66
ViT-S/16					
Multi-run training (18 Models)					
ENS	50.49	30.25	54.44	40.24	43.86
DiWA	52.10	30.70	56.52	41.05	45.09
Single-run training					
ERM	53.66±2.48	27.49±2.25	51.30±0.05	36.75±0.04	42.30±1.20
High-rate Mixout	58.62±1.03	31.71±0.35	56.19±0.60	42.14±0.46	47.16±0.61

Table 6. OOD accuracies on DomainNet.

Method	C	I	P	Q	R	S	**Avg.**
				ResNet50			
Multi-run training (18 Models)							
ENS	68.66	25.39	56.99	14.58	71.28	57.74	49.11
DiWA	66.69	25.15	56.73	14.66	70.40	56.79	48.40
Single-run training							
CORAL	66.89±0.20	24.43±0.20	54.50±0.26	13.82±0.27	68.34±0.31	56.02±0.48	47.33±0.29
Large Dropout	67.04±0.10	25.14±0.28	54.48±0.11	13.37±0.26	68.22±0.15	55.90±0.12	47.36±0.17
ERM	67.09±0.10	25.58±0.32	56.21±0.97	14.85±0.32	69.56±0.52	57.61±0.62	48.48±0.48
High-rate Mixout	66.78±0.30	24.26±0.21	54.90±0.26	14.24±0.28	69.13±0.18	56.84±0.36	47.69±0.26
				ViT-S/16			
Multi-run training (18 Models)							
ENS	68.76	25.82	57.51	16.47	70.18	56.70	49.24
DiWA	66.76	25.87	56.56	16.54	68.62	55.95	48.38
Single-run training							
ERM	66.62±0.10	24.85±0.14	55.04±0.12	15.29±0.35	68.53±0.20	54.60±0.16	47.49±0.18
High-rate Mixout	67.48±0.11	23.68±0.15	54.75±0.19	15.68±0.31	68.16±0.17	54.96±0.15	47.45±0.12