Learning to ignore: Single Source Domain Generalization via Oracle Regularization

Dongkyu Cho[1], Sanghack Lee[1]

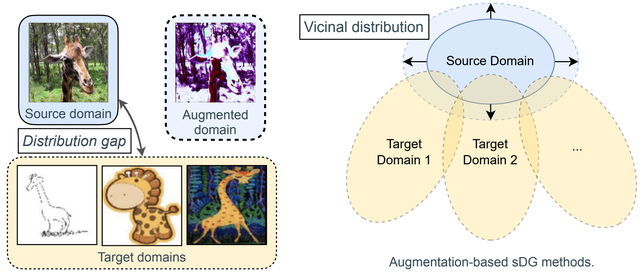
[1] **Seoul National University,** Causality Lab.

kulupapa1127@snu.ac.kr, sanghack@snu.ac.kr

NEURAL INFORMATION PROCESSING SYSTEMS

Problem Setup (sDG)

Single-source Domain Generalization (sDG) is a task designed to simulate domain shift artificially, in order to train a model that can generalize well to multiple unseen target domains from a single source domain. A popular approach is to learn robustness via the alignment of augmented samples.



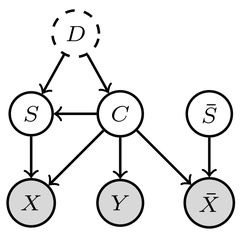
Augmentation-based sDG methods.

Limitations of augmentation-based sDG

Distinction of domain-invariance vs. augmentation-invariance

We highlight overlooked issues in using augmentation for generalization by analyzing the data generating process behind data augmentation.

- X: The observed sample (e.g., Image)
- Y: The associated label
- $\bar{X}$: The augmented view of X
- S: Features changed by augmentation
- C: Features invariant to augmentation
- D: Domain Variable



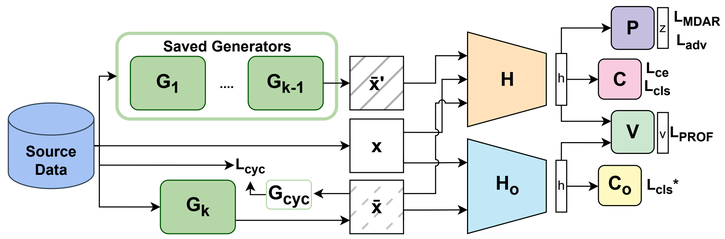
The data generating process

We adopted the causal model from Von Kügelgen et al. [1] and added a new variable: the domain D. It is studied that aligning augmented samples can retrieve the *augment-invariant* features C, which is not necessarily *domain-invariant*, In sDG, we cannot distinguish what information is shared across different domains, leaving both C and S potentially affected by D. Hence, under large domain gaps, augmentation does not guarantee OOD generalization.

We observe a strong correlation between the level of domain gap and the magnitude of performance fluctuation during training. Our hypothesis is that by learning domain-invariant features, we may mitigate the issue.

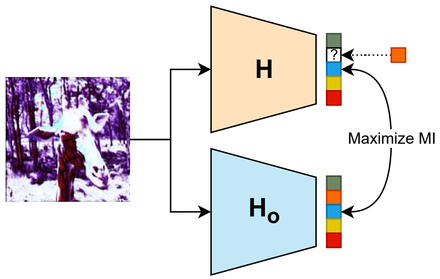
Oracle Regularization: PROF

We present Progressive mutual information Regularization for Online distillation of Frozen oracles, which regulates the learning process.



An overview of our method, PROF.

PROF is an oracle regularization method. The underlying assumption is that a large, pretrained model can approximate an oracle, which can extract domain-invariant features.



Regularization process

- H: Task model
- H_o: Oracle
- Mutual Information(MI):=

$$I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x\,|\,y)}{p(x)}\right]$$

* In practice, we used a variational lower bound of mutual information(MI) for optimization, as direct estimation of MI is difficult.

PROF maximizes the MI between the intermediate output features of the two feature-extractors H (task model) and H_o (Oracle), encouraging the task model to imitate the oracle on what to *ignore* from augmented samples. PROF is defined as:

$$L_{PROF}(x, \bar{x}, \lambda_{PROF}) = \sum_{x' \in \{x, \bar{x}\}} BT(V(H(x')), V(H_o(x')), \lambda_{PROF}),$$

which maximizes the lower bound of MI using a non-contrastive alignment loss (BT), suggested in Zbontar et al. [2]. BT is defined as:

$$\mathrm{BT}(z, z^+, \lambda) = \sum_i (1 - M_{ii})^2 + \lambda \sum_i \sum_{j \neq i} M_{ij}^2.$$

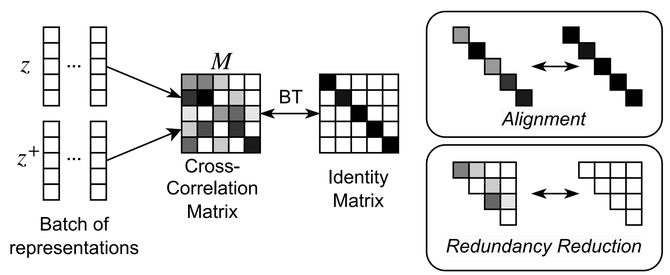


Illustration of the alignment loss (BT; [2]) used in PROF

Experiments

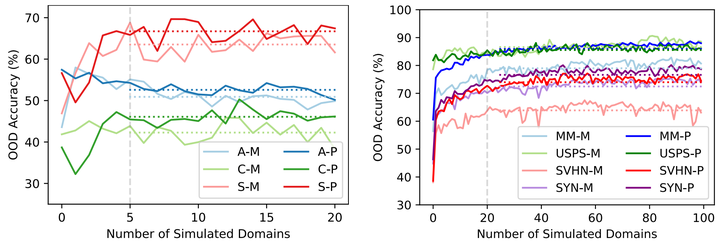
PROF *stabilizes the learning process of augmentation-based sDG.*

PROF effectively reduces the mid-train OOD fluctuation, measured as variance. Our comparative baseline is a conventional augment & align method, designed to work on small batch sizes. Please refer the original paper for details on our baseline.

Method	A	C	S
Baseline (M)	3.39	5.22	7.23
Ours (P)	**1.27**	**2.49**	**5.30**

Method	SVHN	M-M	S-D	USPS
Baseline (M)	3.58	2.56	2.36	3.48
Ours (P)	**1.95**	**1.17**	**2.10**	**1.11**

PROF's effect in stabilizing the mid-train OOD fluctuation, measured as variance. (Left: PACS/ Right: Digits)



PROF's effect on the stabilization of the augmentation-based learning process (Left: PACS/ Right: Digits)

PROF displays competitive generalization scores, using simple augmentation methods compared to SoTA methods.

Method	A	C	S	Avg.
ERM	54.43	42.74	42.02	46.39
ADA	58.72	45.58	48.26	50.85
ME-ADA	58.96	**51.05**	58.42	51.00
L2D (AN)	56.26	51.04	58.42	55.24
MetaCNN	54.05	53.58	63.88	57.17
Baseline (P)	52.46	50.29	**66.79**	56.52
Ours (M)	57.54	46.89	64.93	56.45
Ours (MP)	**58.96**	45.86	64.57	56.46

Method	SVHN	M-M	S-D	USPS	Avg.
ERM	27.83	52.72	39.65	76.94	49.29
JiGen	33.80	57.80	43.79	77.15	53.14
M-ADA	42.55	67.94	48.95	78.53	59.49
L2D	62.86	87.30	63.72	83.97	74.46
PDEN	62.21	82.20	69.39	85.26	74.77
MetaCNN	66.50	**88.27**	70.66	**89.64**	78.76
Baseline (M)	**68.29**	81.88	**76.24**	88.79	**78.80**
Ours (P)	**74.50**	87.98	**78.67**	86.15	**81.82**

PROF's generalization score (Left: PACS/ Right: Digits)

Discussion & Future Works

PROF exploits a large pretrained model for regularization. There are concerns that direct use of the oracle is preferable. Yet, the sDG task requires strict conditions for model architecture (e.g., AlexNet for PACS, 3 layer MLP for Digits), hence we cannot. In the future, however, we aspire to advance our work such that it does not require pretrained oracles.

Furthermore, there are failure cases of PROF. Especially, if the oracle is not able to display domain-invariance, PROF fails to show stabilization effect. For instance, the oracle for PACS cannot be used for corrupted CIFAR-10 experiments, as was is not adversarially-trained.

References

[1] Von Kügelgen et al. (2021). "Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style." In: Advances in neural information processing systems 34 (2021): 16451-16467.

[2] Zbontar et al. (2021) "Barlow twins: Self-supervised learning via redundancy reduction." In: *International Conference on Machine Learning*. PMLR, 2021.

Seoul National University
Graduate School of Data Science