Model Agnostic Sample Reweighting for Domain Generalization

Xiao Zhou*,[1], Yong LIN*,[1], Renjie Pi*,[1], Weizhong Zhang[1], Renjie Xu[2], Peng Cui[2], Tong Zhang[1]

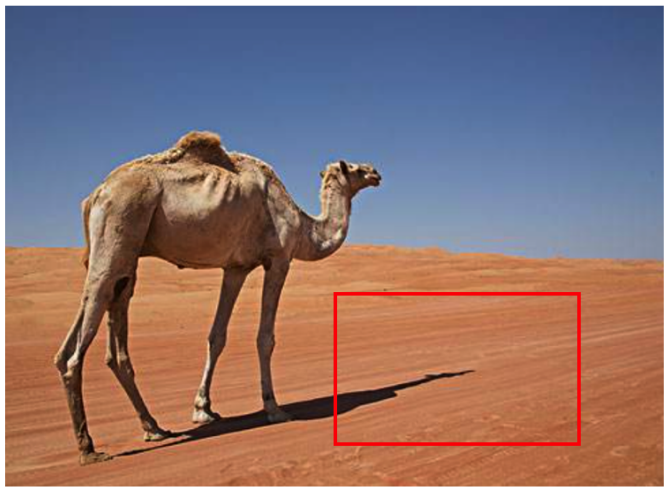
* Equal Contribution, [1]Hong Kong University of Science and Technology, [2]Tsinghua University

July 6, 2022

Background (OOD problem)

Out-of-distribution (OOD) generalization problem:

- The conventional i.i.d. assumption may fail because the testing distribution is the same with the training one.
- This is especially problematic if a model relies on spurious feature which exhibit high correlation with target in the training set.



camel

cow

Background (IRM and DRO)

Invariant Risk Minimization (IRM) [Arjovsky et al., 2019] and Distributional Robust Optimization (DRO) [Sagawa et al., 2019] are two popular methods to alleviate this problem.

$$\mathcal{R}_{\text{IRMv1}}(\mathcal{D}, \boldsymbol{\theta}) := \sum_e \mathcal{L}(\mathcal{D}^e, \boldsymbol{\theta}) + \lambda \|\nabla_v \mathcal{L}(\mathcal{D}^e, \boldsymbol{\theta})\|_2^2 \tag{1}$$

$$\mathcal{R}_{\text{Group-DRO}}(\mathcal{D}, \theta) := \max_e \mathcal{L}(\mathcal{D}^e, \theta) \tag{2}$$

where $\mathcal{L}(\mathcal{D}, \theta)$ is the loss on dataset $\mathcal{D}$ of model θ. However, recent literature shows that IRM and DRO deteriorates dramatically if overfitting occurs, which is commonly the case with large DNN [Lin et al., 2022]. [1]

[1] Yong Lin, et. al., Bayesian Risk Minimization, CVPR 2022

Bilevel Model Agnostic Reweighting (MAPLE)

Motivation:

- Reweighting is a popular technique on mitigating bias (the correlation between Y and spurious feature is a kind of bias).
- If we can find a proper reweighting, we can train a reweighted ERM to learn a invariant feature.

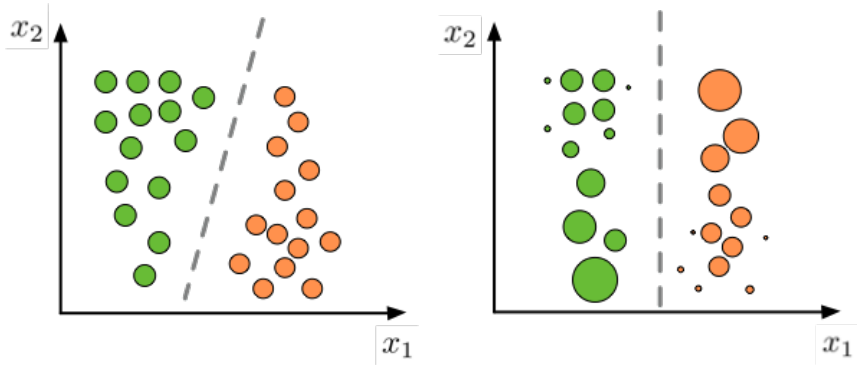


Figure: Left) unweighted; Right) weighted. x_1 and x_2 are the invariant and spurious features, respectively. Fitting a linear classifier $[w_1, w_2]^\top [x_1, x_2]$on unweighted results in a model biased towards x_2 with $w_2 \neq 0$.

MAPLE

We use IRM loss to guide the searching for such weight. The space of sample weights is much smaller than that of the NN parameters.
Consider the reweighting function:

$$\mathcal{S} = \{s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+ | \mathbb{E}[s(\boldsymbol{x}, \boldsymbol{y})] = 1\}.$$

We bilevel method to optimize for the reweighting function:

$$\min_{\boldsymbol{s} \in \mathcal{S}} \mathcal{L}(\boldsymbol{\theta}^*(\boldsymbol{s}); \mathcal{D}_v), \tag{3}$$

$$s.t.\ \boldsymbol{\theta}^*(\boldsymbol{s}) \in \arg\min_{\boldsymbol{\theta}} \mathcal{R}(\boldsymbol{\theta}; \mathcal{D}_{tr}(\boldsymbol{s})), , \tag{4}$$

here $\boldsymbol{\theta} = [w, \Phi]$, $\mathcal{D}_{tr}$ and $\mathcal{D}_v$ are training and validation dataset from the same distribution, respectively. $\mathcal{D}(s)$ is the dataset reweighted by s. $\mathcal{R}(\boldsymbol{\theta}; \mathcal{D})$ and $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$ are the ERM and IRM risk on dataset $\mathcal{D}$.
Specifically:

$$\mathcal{R}(\boldsymbol{\theta}, \mathcal{D}(s)) = \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}} s(\boldsymbol{x}, y) \ell(\boldsymbol{\theta}; \boldsymbol{x}, y)$$

Analysis in a linear case

Consider $\boldsymbol{x} = [\boldsymbol{x}_{inv}, \boldsymbol{x}_s]$. We want to fit a linear model $\boldsymbol{\theta}^\top \boldsymbol{x}$ to predict y.

Lemma (Existence of a "debiased" weighting function)

Given infinite data in the training dataset $\mathcal{D}_{tr}$, there exists a weight function $s \in \mathcal{S}$, i.e.,

$$s(\boldsymbol{x}, y) = \frac{\mathbb{P}(\boldsymbol{x}_{inv}, y)\mathbb{P}(\boldsymbol{x}_s)}{\mathbb{P}(\boldsymbol{x}_{inv}, \boldsymbol{x}_s, y)},$$

such that the solution of Eq. (4) satisfies that

$$\boldsymbol{\theta}^*(s) = \bar{\boldsymbol{\theta}} = [\bar{\boldsymbol{\theta}}_{inv}; \mathbf{0}],$$

where $\bar{\boldsymbol{\theta}}_{inv}$ is the optimal model that merely uses $\boldsymbol{x}_{inv}$, i.e.,

$$\bar{\boldsymbol{\theta}}_{inv} := \underset{\boldsymbol{\theta} \in \mathbb{R}^{d_{inv}}}{\arg\min}\, \mathbb{E}[(y - \boldsymbol{\theta}^\top \boldsymbol{x}_{inv})^2]$$

Analysis in a linear case

Assumption

Given infinite dataset $\mathcal{D}$, the optimal invariant predictor $\bar{\boldsymbol{\theta}}$ is identifiable by the IRM risk $\mathcal{L}$, i.e., $\mathcal{L}(\bar{\boldsymbol{\theta}}, \mathcal{D}) < \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}), \forall \boldsymbol{\theta} \in \mathbb{R}^d, \boldsymbol{\theta} \neq \bar{\boldsymbol{\theta}}$.

This assumption is verified in [Arjovsky et al., 2019] with some conditions.

Theorem (Identifiability of MAPLE)

Assuming infinite data in both $\mathcal{D}_{tr}$ and $\mathcal{D}_v$, when Assumption 1 holds, the populated MAPLE, i.e., Eqn.(3)-(4), can uniquely identify $\bar{\boldsymbol{\theta}}$.

If $\mathcal{D}_{tr}$ and $\mathcal{D}_v$ contain finite samples, we first obtain $\hat{\boldsymbol{\theta}}(s)$ on $\mathcal{D}_{tr}$ by solving Eqn. (4). Regarding $\hat{\boldsymbol{\theta}}(\cdot)$ as a fixed mapping independent of $\mathcal{D}_v$, assuming $\hat{s}$ is a ϵ-approximated solution of MAPLE in. (3), we can also obtain some finite sample properties ($|\mathcal{D}_v| = n$), e.g.,

$$\mathbb{E}[\mathcal{L}(\hat{\boldsymbol{\theta}}(\hat{s}), \mathcal{D}_v)] \leq \min_s \mathbb{E}[\mathcal{L}(\hat{\boldsymbol{\theta}}(s), \mathcal{D}_v)] + \epsilon + C\sqrt{\frac{2\ln(2|\mathcal{S}|/\delta)}{n}}$$

Experiments

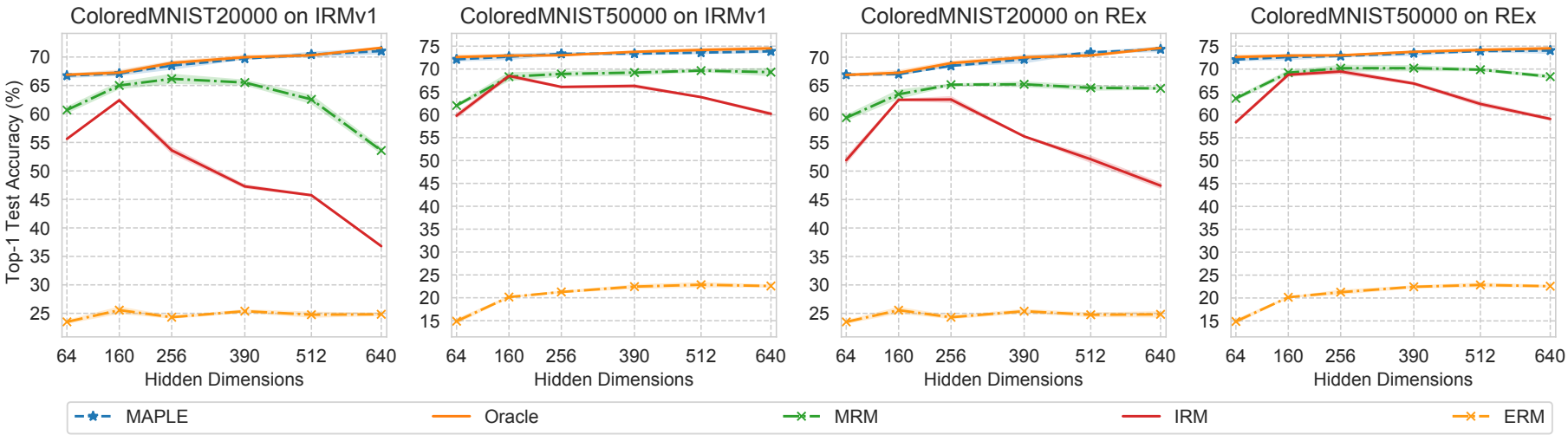


Figure: Comparison between MAPLE and baselines on CMNIST. "Oracle" means training ERM on the dataset without spurious feature, serving as an upper bound.

Experiments

Method	ColoredObject	CifarMnist
Oracle	87.9±0.6	83.7±1.5
ERM	49.8±0.4	39.5±0.4
IRMv1	71.4±0.2	51.3±3.0
REx	73.2±2.9	50.1±2.2
InvRat	73.5±1.5	52.3±0.9
BayesianIRM	78.1±0.6	59.3±2.3
SparseIRM	80.1±1.0	62.3±0.7
MAPLE	**86.9**±0.5	**82.5**±0.4

Table: Test accuracy on IRM tasks with ResNet-18

Experiments

Method	Group Indexes in $\mathcal{D}_{tr}$	Test Average	Test Worst
Upweighting [Cui et al., 2019]	Yes	92.2	87.4
GroupDRO [Sagawa et al., 2019]	Yes	93.5	91.4
ERM	No	97.3	72.6
CVaR DRO [Levy et al., 2020]	No	96.0	75.9
LfF [Nam et al., 2020]	No	91.2	78.0
JTT [Liu et al., 2021]	No	93.3	86.7
MAPLE	No	**92.9**	**91.7**

Table: Comparison of MAPLE and state-of-the-art DRO methods in Waterbirds. The validation set has group annotation following [Liu et al., 2021].

MAPLE

Advantages:

- Mapping the optimization from parameter space to sample weighting space. Alleviating the overfitting problem of IRM (also applicable to DRO).
- Agnostic to the model (the neural network can be easily replaced with another one).

Disadvantages:

- Bilevel training introduces computational overhead, affecting scalability.

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019).
Invariant risk minimization.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019).
Class-balanced loss based on effective number of samples.

Levy, D., Carmon, Y., Duchi, J. C., & Sidford, A. (2020).
Large-scale methods for distributionally robust optimization.

Lin, Y., Dong, H., Wang, H., & Zhang, T. (2022).
Bayesian invariant risk minimization.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16021–16030).

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., & Finn, C. (2021).
Just train twice: Improving group robustness without training group information.
In *International Conference on Machine Learning* (pp. 6781–6792).: PMLR.

Nam, J., Cha, H., Ahn, S., Lee, J., & Shin, J. (2020).

Learning from failure: Training debiased classifier from biased classifier.
arXiv preprint arXiv:2007.02561.

Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.
arXiv preprint arXiv:1911.08731.