UMIX: Improving Importance Weighting for Subpopulation Shift via Uncertainty-Aware Mixup

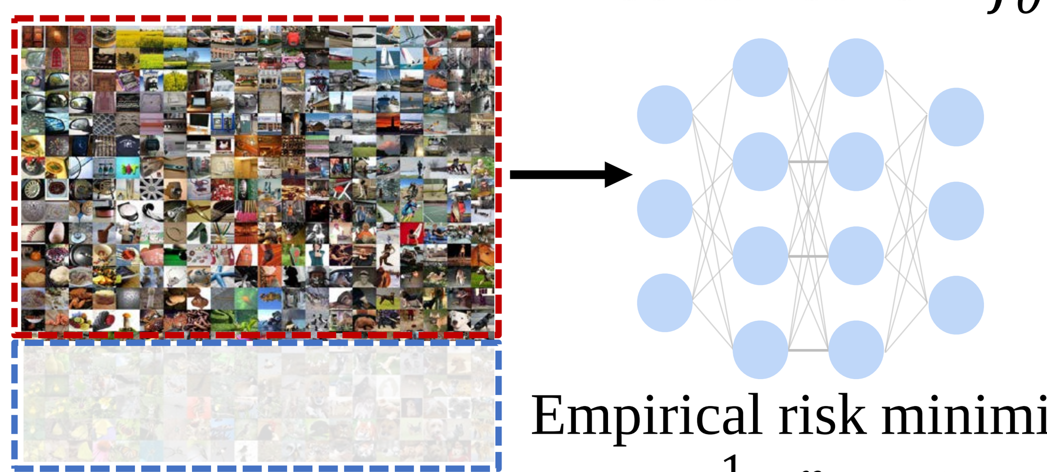
Zongbo Han, Zhipeng Liang, Fan Yang, Liu Liu, Lanqing Li, Yatao Bian, Peilin Zhao, Bingzhe Wu, Changqing Zhang, Jianhua Yao

Tianjin University, Hong Kong University of Science and Technology, Tencent AI Lab

(1) Machine Learning Paradigm

I.I.D. Assumption: $P_{tr}(X,Y) = P_{te}(X,Y)$.

Neural network f_θ



Empirical risk minimization

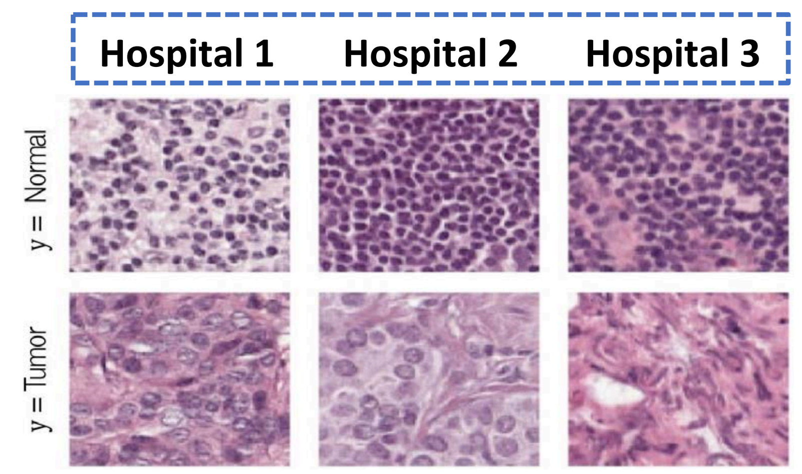
$$\mathcal{L}_{ERM} = \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)$$

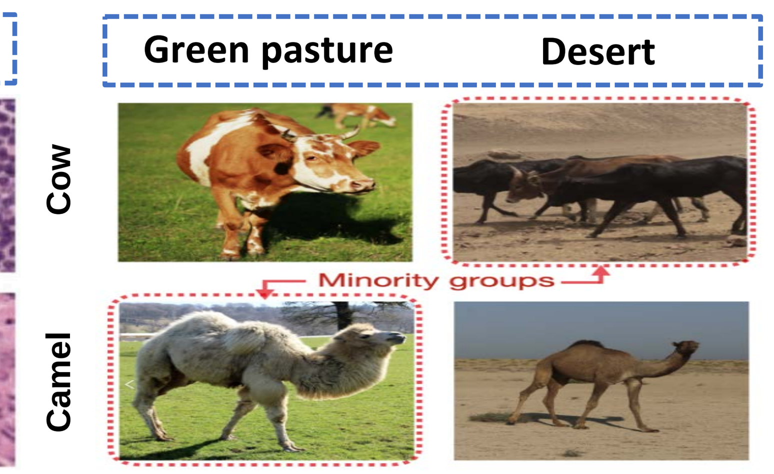
ERM faces challenges from distribution shift

(2) Definition of Subpopulation Shift

$P_{tr}(X,Y)$ is a mixture of G predefined subpopulations, i.e., $P_{tr} = \sum_{g=1}^{G} k_g P_g$.

Key Definition: $P_{tr} = \sum_{g=1}^{G} k_g P_g$





(3) ERM and Importance weighting

ERM $\mathbb{E}_{(x,y)\sim P_{tr}} \ell(f_\theta(x), y) = \boxed{\sum_{g=1}^{G} k_g} \mathbb{E}_{(x,y)\sim P_g} \ell(f_\theta(x), y)$

The model tends to *focus on the majority subpopulations* **in the training set.**

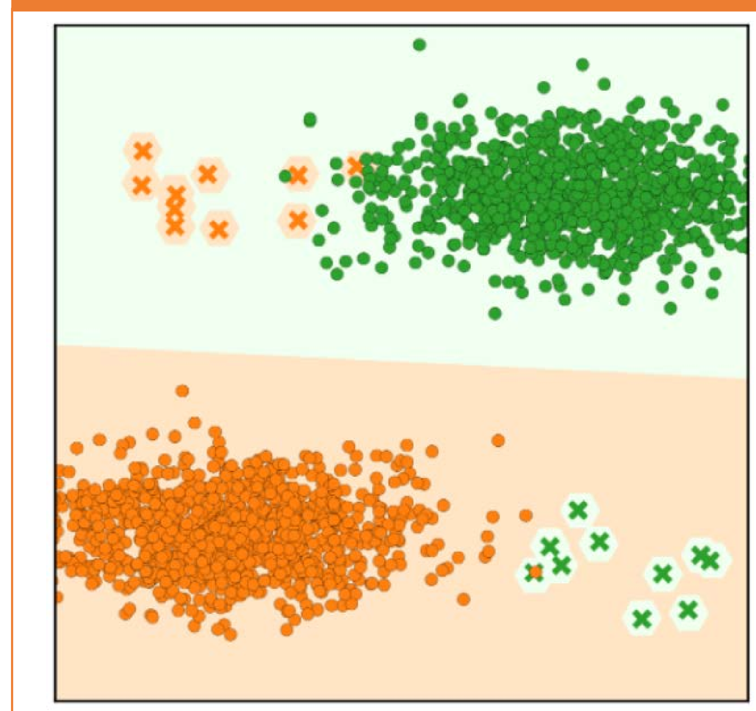
Our objective: find an optimal model f_θ^* which can **generalize best on the worst-case** subpopulation data:

$$f_\theta^* = \operatorname{argmin}_{f_\theta} \max_{g=1,\dots,G} \mathbb{E}_{x,y\sim P_g}[\ell(f_\theta(x), y)].$$

Weighted ERM $\mathbb{E}_{(x,y)\sim P_{tr}} w(x,y) \ell(f_\theta(x), y)$

imposing static or adaptive weight on each sample and then building weighted empirical loss. Therefore each subpopulation group can have a comparable strength in the final training objective.

(4) IW for Overparameterized NNs



Recent studies have shown both empirically and theoretically that importance weighting methods could fail to achieve better worst-case subpopulation performance especially when they are applied to over-parameterized neural networks (NNs).

Overparameterized neural networks memorize the minority points[1].

(5) Importance-weighted mixup

Linear interpolations:

$$\tilde{x}_{i,j} = \lambda x_i + (1-\lambda)x_j, \ \tilde{y}_{i,j} = \lambda y_i + (1-\lambda) y_j$$

Loss function:

- Vanilla mixup $\mathbb{E}[\lambda\ell(\theta, \tilde{x}_{i,j}, y_i) + (1-\lambda)\ell(\theta, \tilde{x}_{i,j}, y_j)]$
- Ours (UMIX) $\mathbb{E}[w_i \lambda\ell(\theta, \tilde{x}_{i,j}, y_i) + w_j(1-\lambda)\ell(\theta, \tilde{x}_{i,j}, y_j)]$

Uncertainty-aware importance weights

The proposed method combines the advantages of MIXUP and importance reweighting.

(6) importance weights

During training, easy samples are easier to learn, while hard samples are more likely to be misclassified[2].

$u_i \approx \frac{1}{T}\sum_{t=T_s}^{T_s+T} \kappa(y_i, \hat{f}_{\theta_t}(x_i))$ $\quad w_i = \eta u_i + c$

(7) A tighter generalization bound

Theorem 5.1. *Suppose $A(\cdot)$ is L_A-Lipschitz continuous, then there exists constants $L, B > 0$ such that for any θ satisfying $\theta^\top \Sigma_X \theta \le \gamma$, the following holds with a probability of at least $1-\delta$,*

$$\text{GError}(\theta) \le 2L \cdot L_A \cdot (\max\{(\frac{\gamma(\delta/2)}{\rho})^{1/4}, (\frac{\gamma(\delta/2)}{\rho})^{1/2}\} \cdot \boxed{\sqrt{\frac{\text{rank}(\Sigma_X)}{n}})} + B\sqrt{\frac{\log(2/\delta)}{2n}},$$

where $\gamma(\delta)$ is a constant dependent on δ and $\Sigma_X = \sum_{g=1}^{G} k_g w_g \Sigma_X^g$.

Weighted ERM	Ours
$\sqrt{\frac{d}{n}}$	$\sqrt{\frac{\text{rank}(\sum_{g=1}^{G} k_g w_g \Sigma_X^g)}{n}}$

$$\text{rank}(\Sigma_X) \ll d$$

In contrast to weighted ERM, the bound improvement of UMIX is on the red term which can partially reflect the heterogeneity of the training subpopulations.

(8) Mainly Experimental results

	Waterbirds		CelebA		CivilComments	
	Avg.	Worst	Avg.	Worst	Avg.	Worst
ERM	**97.0%**	63.7%	**94.9%**	47.8%	**92.2%**	56.0%
Focal Loss [34]	87.0%	73.1%	88.4%	72.1%	91.2%	60.1%
CVaR-DRO [32]	90.3%	77.2%	86.8%	76.9%	89.1%	62.3%
CVaR-DORO [63]	91.5%	77.0%	89.6%	75.6%	90.0%	64.1%
χ^2-DRO [32]	88.8%	74.0%	87.7%	78.4%	89.4%	64.2%
χ^2-DORO [63]	89.5%	76.0%	87.0%	75.6%	90.1%	63.8%
JTT [35]	93.6%	86.0%	88.0%	81.1%	90.7%	67.4%
Ours	93.0%	**90.0%**	90.1%	**85.3%**	90.6%	**70.1%**

	Group labels in train set?	Waterbirds		CelebA		CivilComments	
		Avg.	Worst	Avg.	Worst	Avg.	Worst
IRM [3]	Yes	87.5%	75.6%	**94.0%**	77.8%	88.8%	66.3%
IB-IRM [1]	Yes	88.5%	76.5%	**93.6%**	85.0%	89.1%	65.3%
V-REx [28]	Yes	88.0%	73.6%	92.2%	**86.7%**	**90.2%**	64.9%
CORAL [33]	Yes	90.3%	79.8%	**93.8%**	76.9%	88.7%	65.6%
GroupDRO [51]	Yes	**91.8%**	**90.6%**	92.1%	**87.2%**	89.9%	70.0%
DomainMix [61]	Yes	76.4%	53.0%	93.4%	65.6%	**90.9%**	63.6%
Fish [53]	Yes	85.6%	64.0%	93.1%	61.2%	89.8%	**71.1%**
LISA [62]	Yes	**91.8%**	**89.2%**	92.4%	**89.3%**	89.2%	**72.6%**
Ours	**No**	**93.0%**	**90.0%**	90.1%	85.3%	**90.6%**	**70.1%**

[1] Sagawa S, Raghunathan A, Koh P W, et al. An investigation of why overparameterization exacerbates spurious correlations[C]// ICML 2020

[2] Moon J, Kim J, Shin Y, et al. Confidence-aware learning for deep neural networks[C]//ICML 2020

NEURAL INFORMATION PROCESSING SYSTEMS