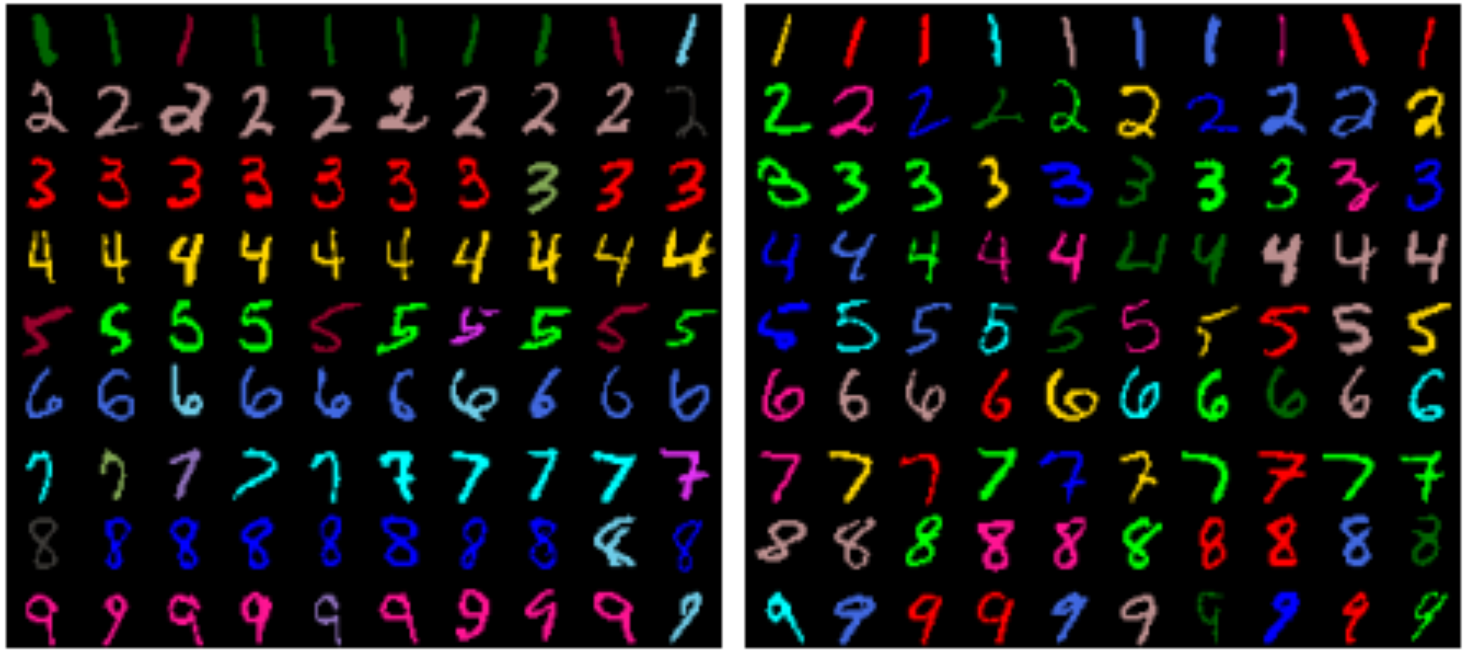
Intro of Out-of-distribution Generalization

Dinghuai Zhang 2020.12

What’s “spurious” correlation?

“true label” and “spurious label”

What’s “spurious” correlation?



“true label” and “spurious label”

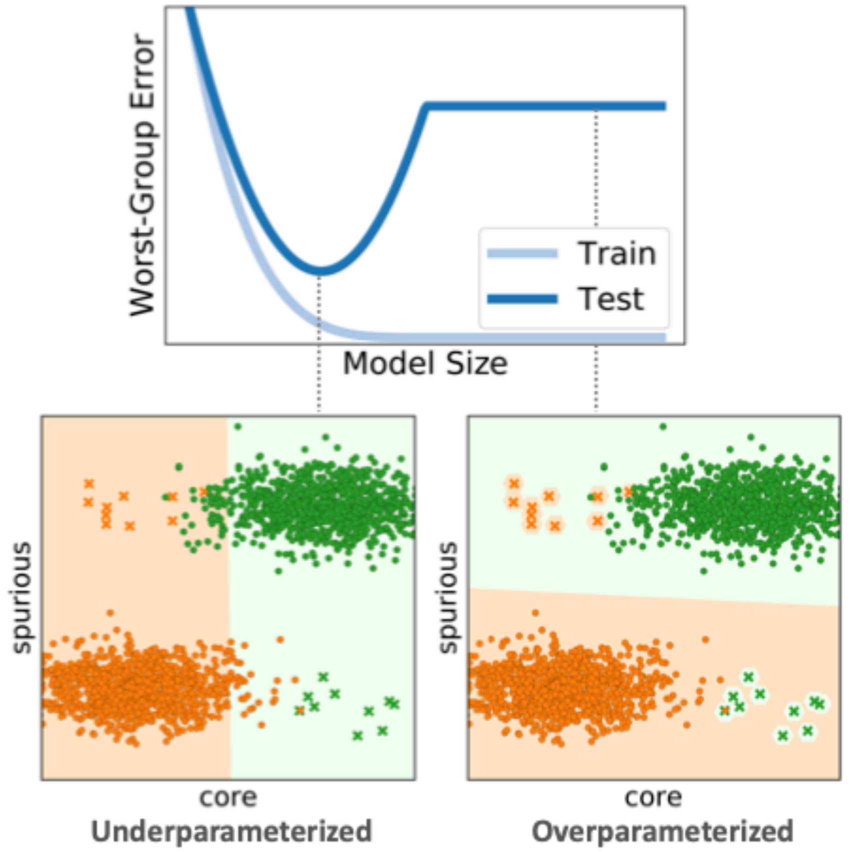
GroupDRO

$$\hat{\theta}_{\mathrm{DRO}} := \underset{\theta \in \Theta}{\arg\min} \Big\{ \hat{\mathcal{R}}(\theta) := \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y) \sim \hat{P}_g} [\ell(\theta; (x, y))] \Big\},$$

DISTRIBUTIONALLY ROBUST NEURAL NETWORKS FOR GROUP SHIFTS:
ON THE IMPORTANCE OF REGULARIZATION FOR WORST-CASE GENERALIZATION, Shiori Sagawa et al.

			Average Accuracy		Worst-Group Accuracy	
			ERM	DRO	ERM	DRO
Standard Regularization	Waterbirds	Train	100.0	100.0	100.0	100.0
		Test	97.3	97.4	60.0	76.9
	CelebA	Train	100.0	100.0	99.9	100.0
		Test	94.8	94.7	41.1	41.1
	MultiNLI	Train	99.9	99.3	99.9	99.0
		Test	82.5	82.0	65.7	66.4
Strong ℓ_2 Penalty	Waterbirds	Train	97.6	99.1	35.7	97.5
		Test	95.7	96.6	21.3	84.6
	CelebA	Train	95.7	95.0	40.4	93.4
		Test	95.8	93.5	37.8	86.7

Overparameterization exacerbates spurious correlations

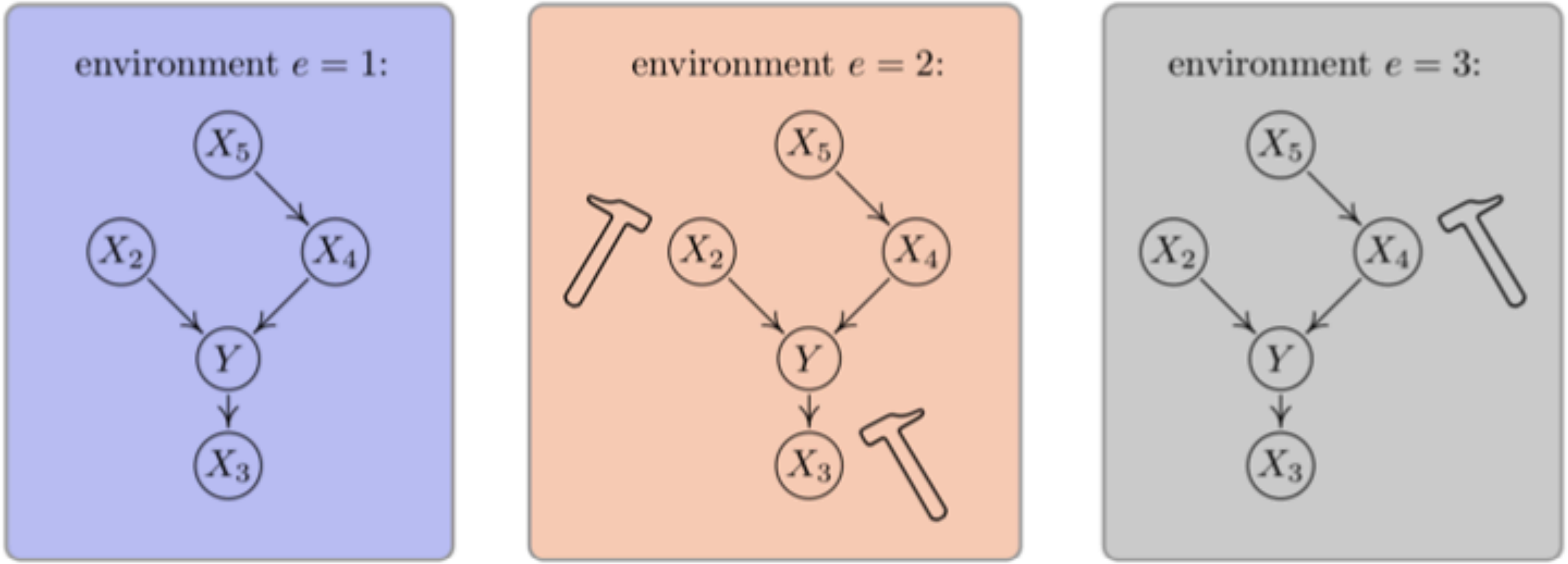


As model size grows, avg errors decrease, but worst group error increases

Reason: overparametrized models use spurious feature to classify

An investigation of why overparameterization exacerbates spurious correlations, Shiori Sagawa et al.

How to get rid of "spurious" feature?
Or, how to do invariant learning



Causal inference using invariant prediction: identification and confidence intervals. Jonas Peters et al

Invariant Causal Predictoin (ICP)

Assumption 1 (Invariant prediction) *There exists a vector of coefficients* $\gamma^* = (\gamma_1^*, \ldots, \gamma_p^*)^t$ *with support* $S^* := \{k : \gamma_k^* \neq 0\} \subseteq \{1, \ldots, p\}$ *that satisfies*

$$\text{for all } e \in \mathcal{E}: \quad X^e \text{ has an arbitrary distribution and}$$
$$Y^e = \mu + X^e\gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}, \tag{3}$$

where $\mu \in \mathbb{R}$ *is an intercept term,* ε^e *is random noise with mean zero, finite variance and the same distribution* F_ε *across all* $e \in \mathcal{E}$.

We will interchangeably use “domain” and “environment”.

Causal Transfer Learning

Algorithm 1: Subset search

Inputs: Sample $(\mathbf{x}_i^k, y_i^k)_{i=1}^{n_k}$ for tasks $k \in \{1, \ldots, D\}$, threshold δ for independence test.

Outputs: Estimated invariant subset $\hat{S}$.

```
Set S_acc = {}, MSE = {}.
for S ⊆ {1,...,p} do
    linearly regress Y on X_S and compute the residuals R_{β^{CS(S)}} on a validation set.
    compute H = HSIC_b((R_{β^{CS(S)},i}, K_i)_{i=1}^n) and the corresponding p-value p* (or
      the p-value from an alternative test, e.g., Levene test.).
    if p* > δ then
        compute Ê_{P1,...,D}(β^{CS(S)}), the empirical estimate of E_{P1,...,D}(β^{CS(S)}) on a
          validation set.
        S_acc.add(S), MSE.add(Ê_{P1,...,D}(β^{CS(S)}))
    end
end
Select Ŝ according to RULE, see Section 3.4.
```

Invariant Models for Causal Transfer Learning Mateo Rojas-Carulla et al

Invariant Risk Minimization

$$\begin{aligned} \min_{\substack{\Phi:\mathcal{X}\to\mathcal{H} \\ w:\mathcal{H}\to\mathcal{Y}}} \quad & \sum_{e\in\mathcal{E}_{\mathrm{tr}}} R^e(w\circ\Phi) \\ \text{subject to} \quad & w \in \underset{\bar{w}:\mathcal{H}\to\mathcal{Y}}{\arg\min}\, R^e(\bar{w}\circ\Phi), \text{ for all } e\in\mathcal{E}_{\mathrm{tr}}. \end{aligned} \tag{IRM}$$

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} R^e(\Phi) + \lambda\cdot \left\|\nabla_{w|w=1.0}\, R^e(w\cdot\Phi)\right\|^2, \tag{IRMv1}$$

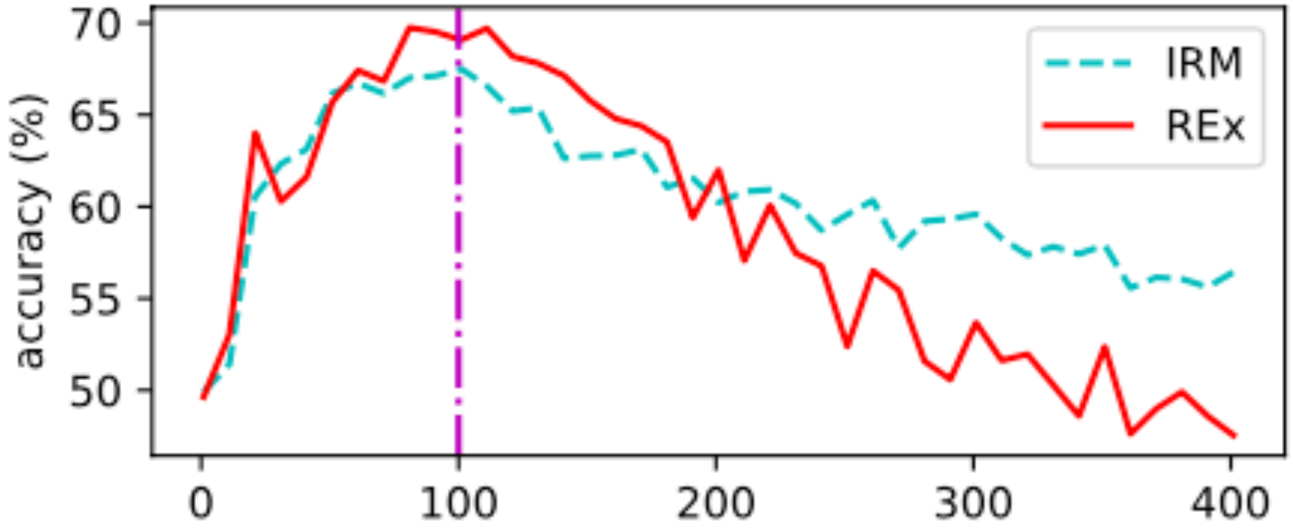
ColoredMNIST

- Binary classification: 0~4 as positive class, 5~9 as negative class
- Each image is either red or green
- Domain1 (train): In all positive images, 70% are red; in all negative images, 30% are red.

Algorithm	Acc. train envs.	Acc. test env.
ERM	87.4 ± 0.2	17.1 ± 0.6
IRM (ours)	70.8 ± 0.9	$\mathbf{66.9 \pm 2.5}$
Random guessing (hypothetical)	50	50
Optimal invariant model (hypothetical)	75	75
ERM, grayscale model (oracle)	73.5 ± 0.2	73.0 ± 0.4

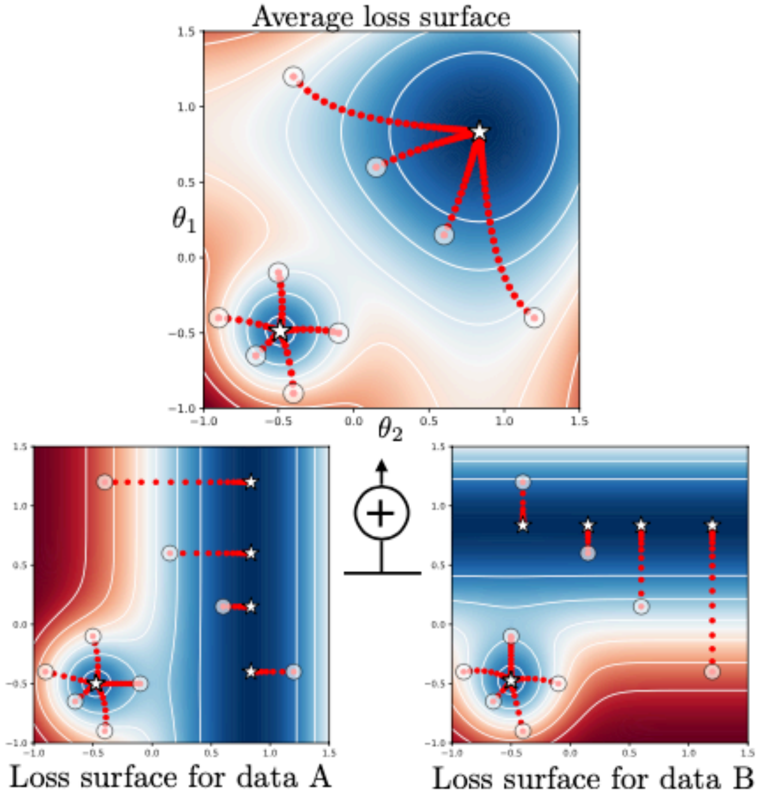
REx

$$\mathcal{R}_{\text{V-REx}} \doteq \beta \mathrm{Var}(\{\mathcal{R}_1, ..., \mathcal{R}_m\}) + \sum_{e=1}^{m} \mathcal{R}_e$$



Out-of-Distribution Generalization via Risk Extrapolation (REx), David Krueger et al.

Learning explanations that are hard to vary



$$\mathcal{C}^{\epsilon}(\theta^*) := - \max_{(e,e')\in\mathcal{E}^2} \max_{\theta \in N^{\epsilon}_{e,\theta^*}} |\mathcal{L}_{e'}(\theta) - \mathcal{L}_e(\theta)|.$$

Learning explanations that are hard to vary Giambattista Parascandolo et al.

“and mask”

a *threshold* $\tau \in [0, 1]$

$$[m_\tau]_j = \mathbb{1}\left[\tau d \leq \left|\sum_e \text{sign}([\nabla \mathcal{L}_e]_j)\right|\right]$$

$$m_t(\theta) \odot \nabla \mathcal{L}(\theta)$$

Learning explanations that are hard to vary Giambattista Parascandolo et al.

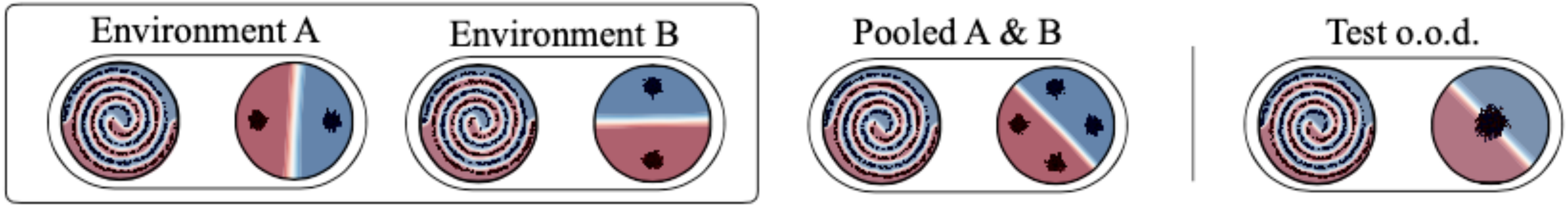


Figure 5: A 4-dimensional instantiation of the synthetic memorization dataset for visualization. Every example is a dot in both circles, and it can be classified by finding either of the "oracle" decision boundaries shown.

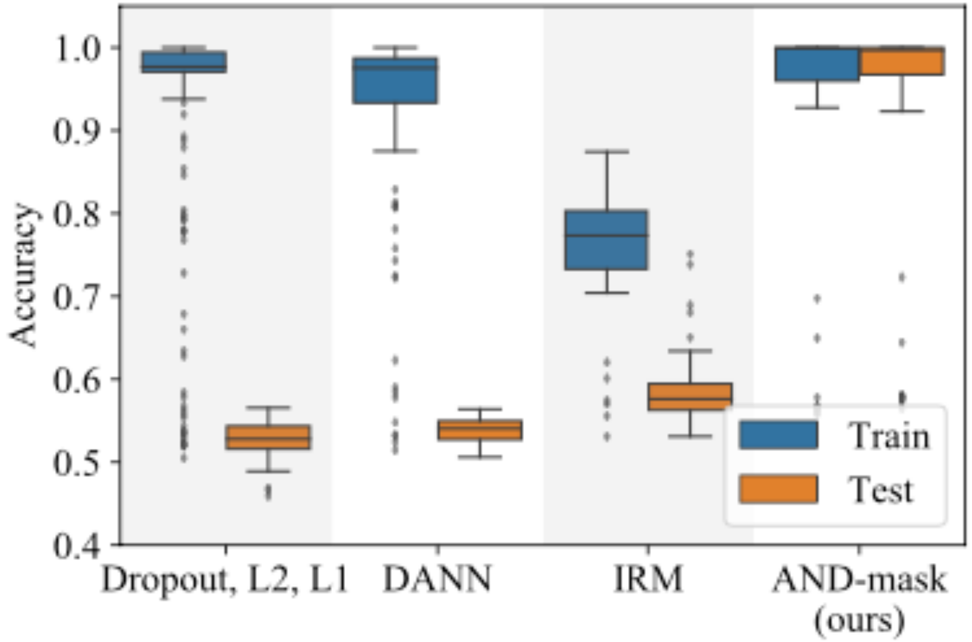


Figure 6: Results on the synthetic dataset.

CIFAR10 random label

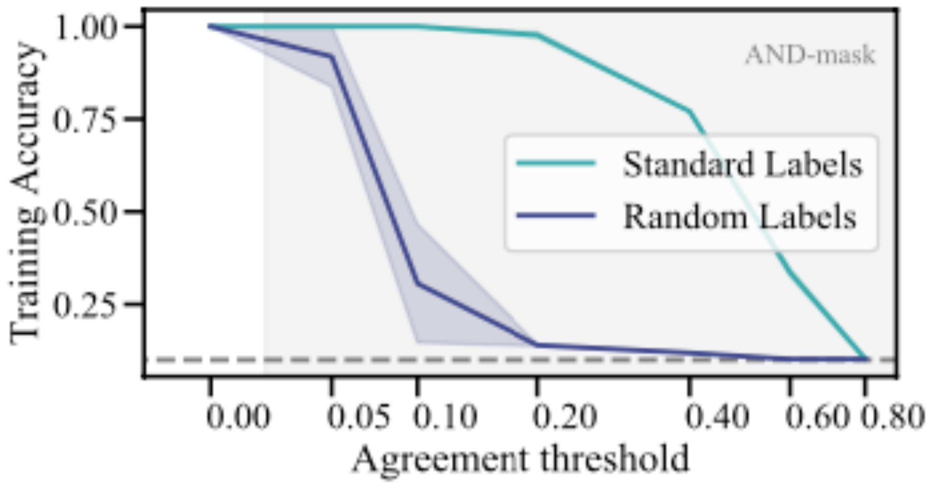
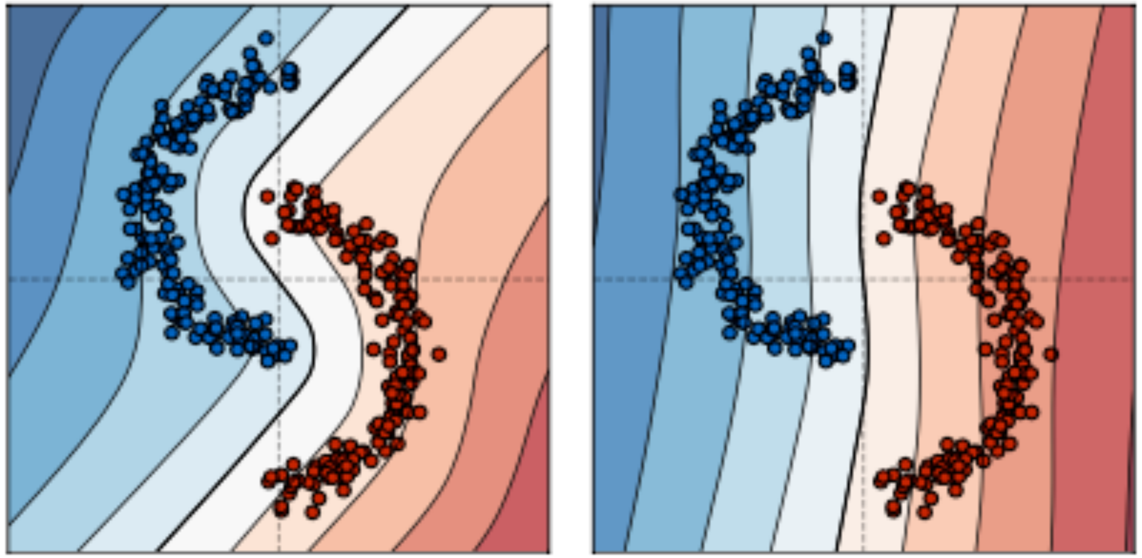


Figure 8: As the AND-mask threshold increases, memorization on CIFAR-10 with random labels is quickly hindered.

Learning explanations that are hard to vary Giambattista Parascandolo et al.

Gradient Starvation

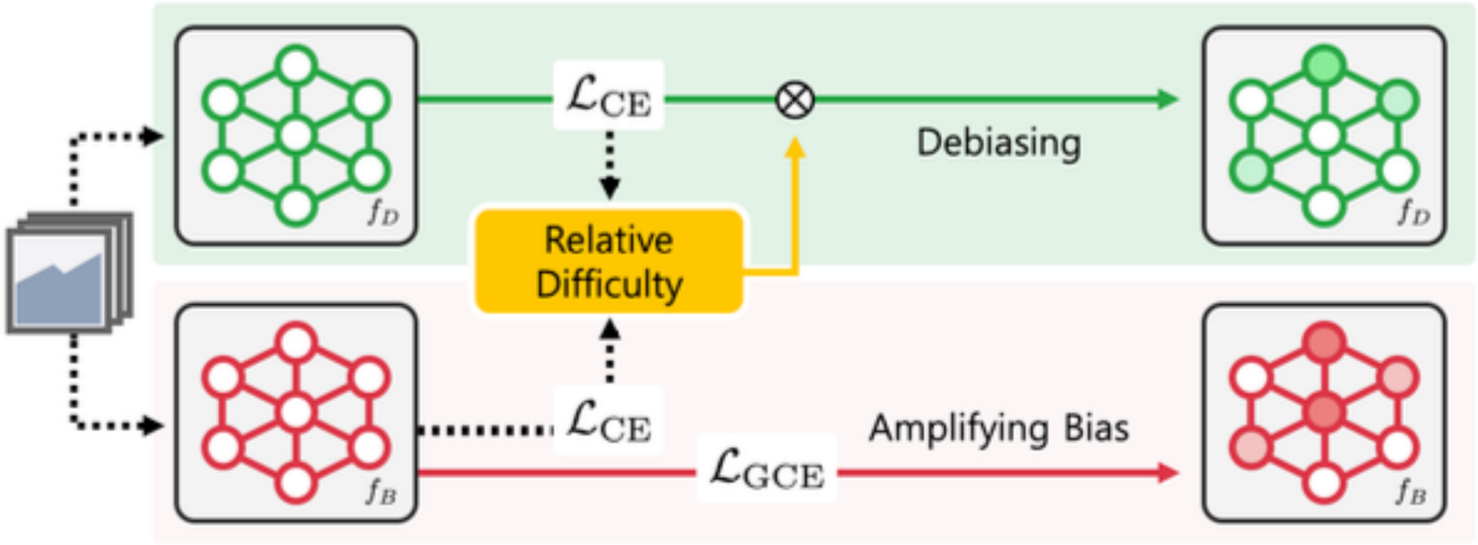


“overfitting” property of ERM

Gradient Starvation: A Learning Proclivity in Neural Networks Mohammad Pezeshki et al.

Learning from Failure

- Setting: eg. No multiple domains
- 99% data: label & color has 1 to 1 corresponding
- 1% data: label & color has no corresponding



Learning from Failure: Training Debiased Classifier from Biased Classifier Junhyun Nam et al.

- More papers at https://sites.google.com/site/irinarish/ood_generalization
- Thank you very much!