

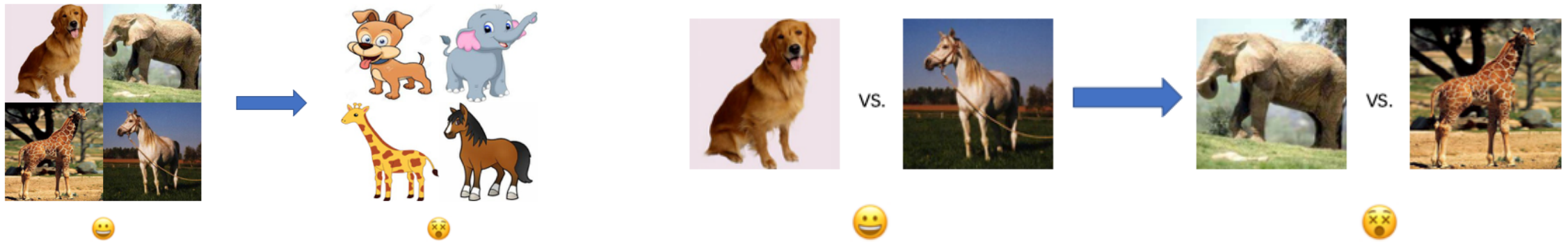
# Can Subnetwork Structure be the Key to Out-of-Distribution Generalization?

Dinghuai Zhang, Kartik Ahuja, Yilun Xu, Yisen Wang, Aaron Courville



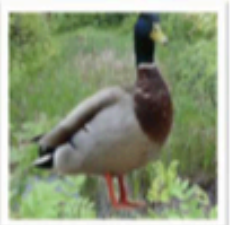

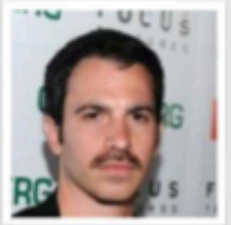

Mila, MIT, Peking University

# Distribution shift problems

- Generalization is one of the core problems in machine learning
- Deep learning has addressed IID generalization to a large extent
- But out-of-distribution (OOD) generalization problem is still far from cooked



# Spurious correlation

	Common training examples		Test examples			
<b>Waterbirds</b>	y: waterbird a: water background		y: landbird a: land background		y: waterbird a: land background	
<b>CelebA</b>	y: blond hair a: female		y: dark hair a: male		y: blond hair a: male	

“True label” may be spuriously correlated with some spurious properties

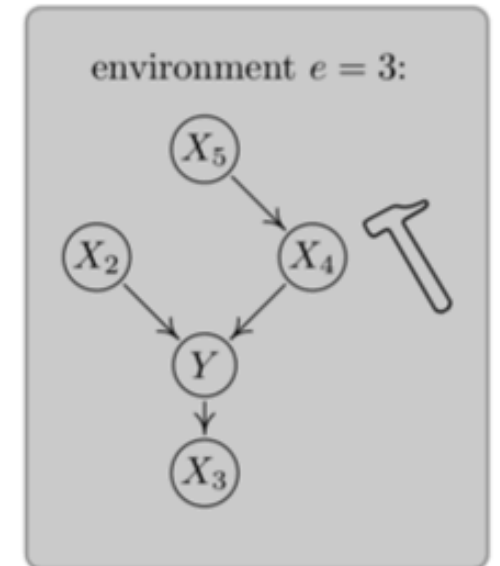
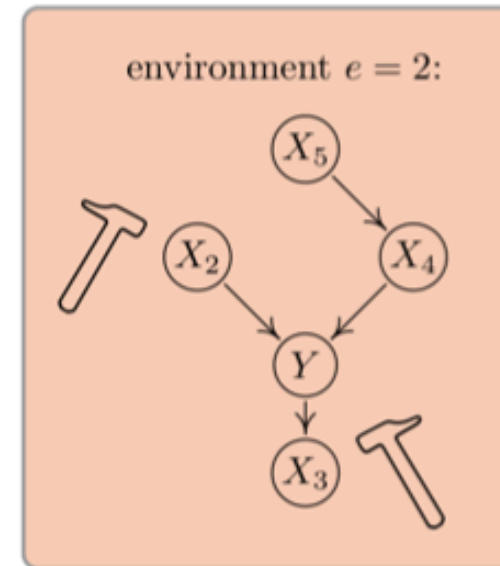
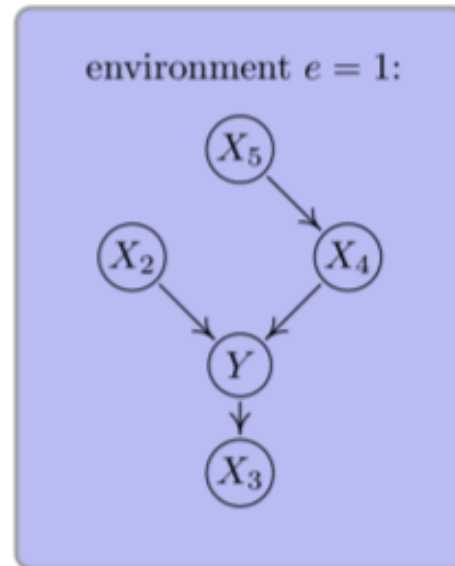
# Multiple environments

“Data is collected in different settings and then shuffled to serve as ‘IID’ ones.”

“Shuffling data is a loss of information.”

--- Leon Bottou

Data collected from different environments follow different distributions



# Out-of-distribution generalization problem

Consider a supervised learning setting where data follows  $(X^e, Y^e) \sim \mathbb{P}^e$

Multiple environments assumption: each environment  $e \in \mathcal{E} = \{1, \dots, E\}$

We only have a subset of environments in training time  $\mathcal{E} = \mathcal{E}_{\text{seen}} \cup \mathcal{E}_{\text{unseen}}$

The goal of OOD generalization problem is defined as  $\min_{\theta \in \Theta} \max_{e \in \mathcal{E}} \mathcal{R}^e(\theta)$

# Related works

Many works from computer vision community (domain adaptation / generalization, transfer learning) propose to match some kind of feature across domains:

match  $P(\Phi(X))$ : Domain-Adversarial Training of Neural Networks, etc.  
match  $P(\Phi(X) | Y)$ : Conditional Adversarial Domain Adaptation, etc.

To the contrary, IRM (Arjovsky et al. 2019) proposes to match  $P(Y | \Phi(X))$  to learn a “causal invariant mechanism” from data. Many follow-ups are proposed to pursue this target ...

# Data structure

We assume input data  $X^e$  is generated from  $Z^e = (Z_{\text{inv}}^e, Z_{\text{sp}}^e)$

$$X^e = G(Z_{\text{inv}}^e, Z_{\text{sp}}^e)$$

We follow the “realizable” assumption<sup>1</sup>, where  $Y_e = F(Z_{\text{inv}}^e)$

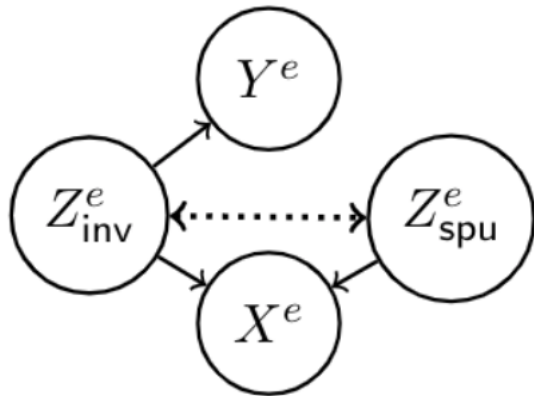
We also assume there exists inverse maps  $Z_{\text{inv}}^e = G_{\text{inv}}^\dagger(X^e)$   $Z_{\text{sp}}^e = G_{\text{sp}}^\dagger(X^e)$

The goal is then to learn  $F \circ G_{\text{inv}}^\dagger$

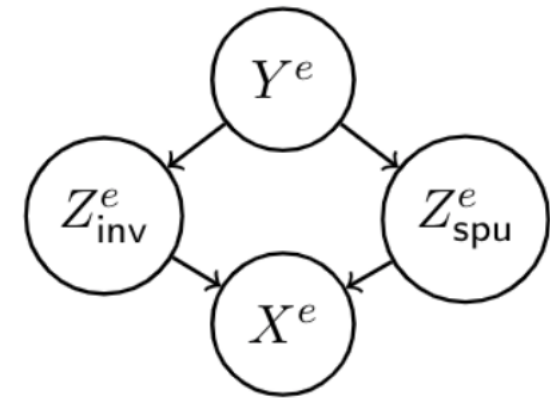
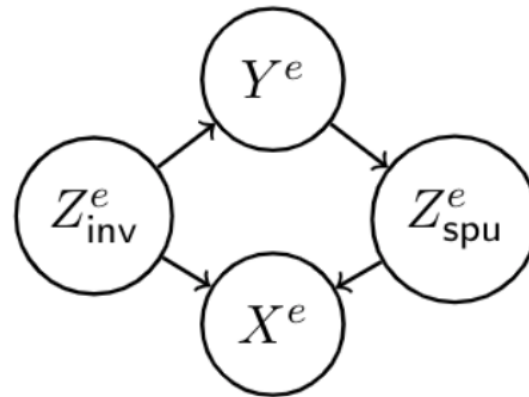
# Other settings

“Realizability” means invariant features contain all information about label

In this work we consider  
“realizable” case



But “non realizable” cases are also possible:



Invariant Risk Minimization, arxiv 2019

Risks of Invariant Risk Minimization, ICLR2021



# A (linear) motivating example

Suppose all labels and latent features are binary

Bias in data: let  $Z_{\text{sp}}^e$  and  $Y^e$  have a  $p^e$  correlation

When  $Z_{\text{sp}}^e$  is high dimensional, the model tends to rely on it<sup>1</sup>

*Proposition (informal):* for a sparse classifier  $f_{\text{sparse}}^d$  and regular classifier  $f_{\text{reg}}$  on dataset with such bias, when the dimensionality of spurious feature is large enough:

- $f_{\text{sparse}}^d$  and  $f_{\text{reg}}$  have similar in-distribution performance
- $f_{\text{sparse}}^d$  has better margin and out-of-distribution performance

# Insights

- Sparsity on proper places has good inductive bias for OOD generalization
- In Peters et al (2015), this is also the case where the proposed algorithm only use subset of linear features, corresponding to sparsity in parameters
- How should we push this insight into deep neural networks?

# Functional Modularity Analysis

A neural network  $f(\mathbf{w}_1, \dots, \mathbf{w}_L; \cdot)$  is parametrized by  $\theta = \{\mathbf{w}_1, \dots, \mathbf{w}_L\}$

We search for a module / subnetwork  $f(\mathbf{m}_1 \odot \mathbf{w}_1, \dots, \mathbf{m}_L \odot \mathbf{w}_L; \cdot)$   
with module mask  $\mathbf{m}_l \in \{0, 1\}^{n_l}$

The subnetwork structure is learned end-to-end with Gumbel-sigmoid trick

Four algorithms are studied: ERM, IRM, REx, GroupDRO

Invariant Risk Minimization, arxiv 2019

Out-of-Distribution Generalization via Risk Extrapolation, ICML2021

Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization, ICLR2020

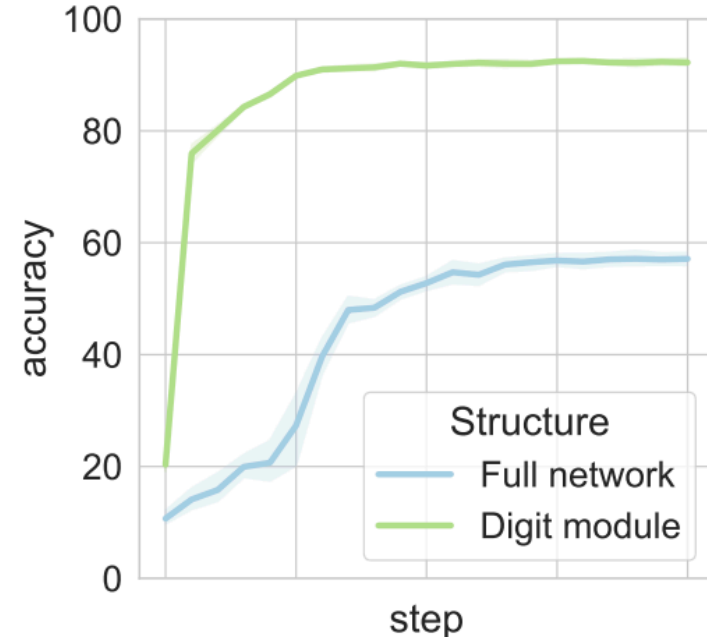
# Modular subnetwork introspection

Does a good subnetwork for OOD exist within a spuriously biased large network?

We construct a 10-class “FullColoredMNIST” with previous stated bias for modularity probing, where digit is invariant feature

We use data which has the same distribution with out-domain to search for a digit module which is good for OOD

We take ERM as an example here:



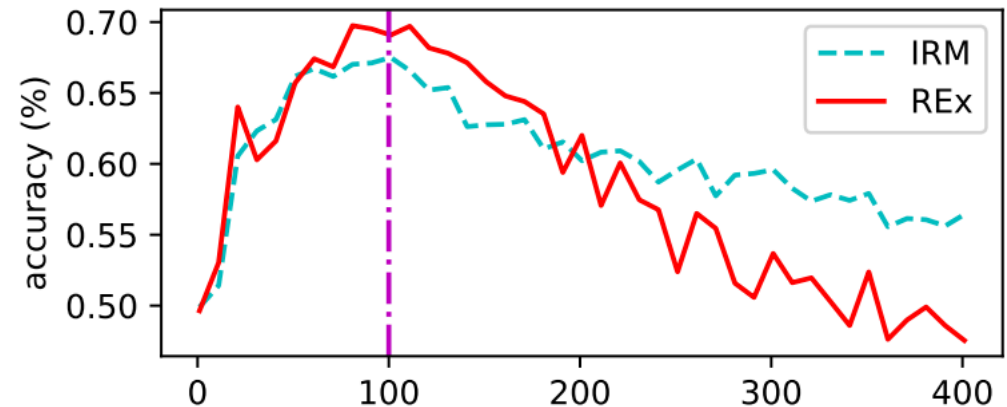
# Feature selection viewpoint

A subnetwork with good inductive bias exists within a large network!

This partially reflect that invariant feature could be extracted effectively with proper structure.

In fact, people have claimed the OOD algorithms are doing feature selection:

(OOD penalty terms can only works well when they're involved at 100 steps)



# Functional “lottery ticket”

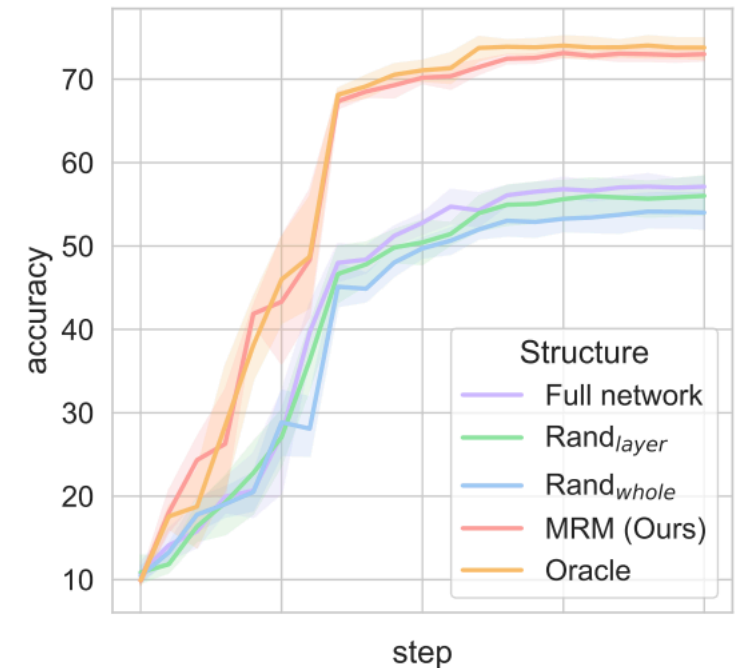
Frankle et al. proposes that there exists subnetwork good for IID generalization

We show that a functional variant of it exists for OOD settings

We propose Modular Risk Minimization (MRM), a straight forward yet effective method to find a good OOD module:

1. train the full model
2. searching module with some desired OOD & sparsity properties
3. retrain the module with same initialization

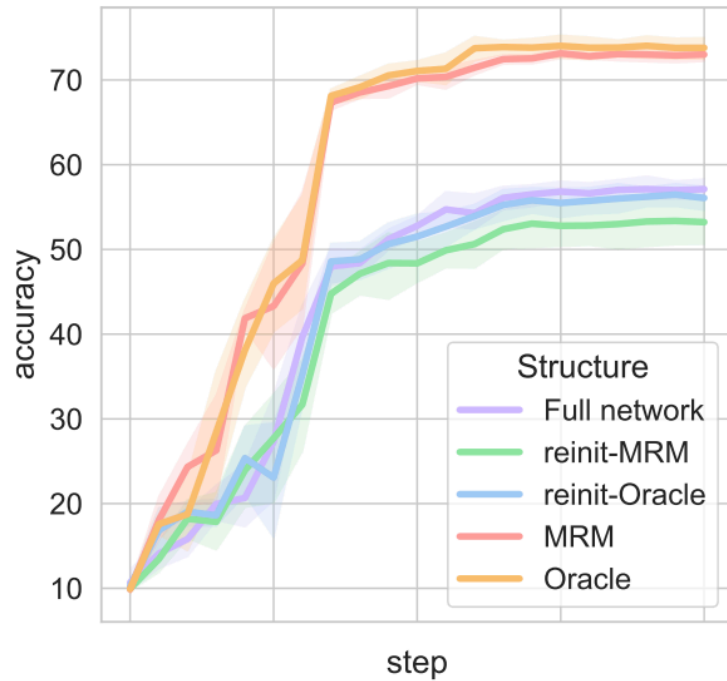
MRM is designed to be easy to combine with other invariant methods like IRM, REx, ... and becomes ModIRM, ModREx, ... See the full paper for more details



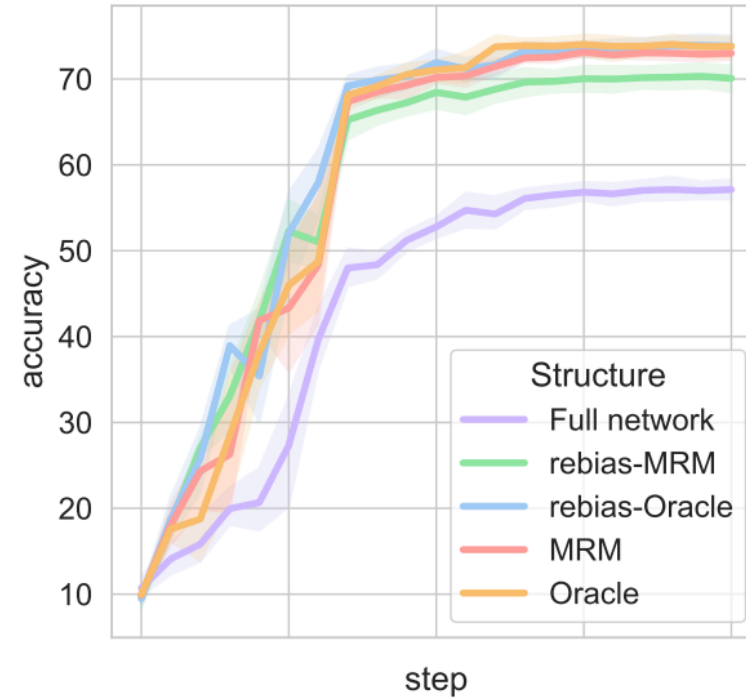
(Oracle means searching module with extra information about test domain in step 2)

# Ablation study

## 1. Importance of initialization



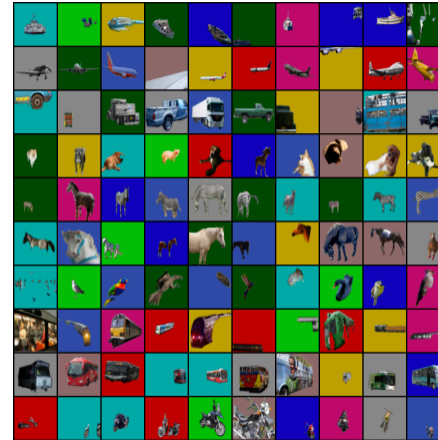
## 2. Effects of bias relationship



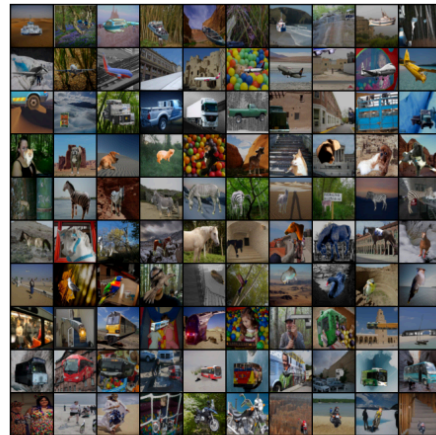
The learned structure is invariant to any kinds of spurious color relationship!

# More experiments

METHODS	TRAIN ACCURACY	TEST ACCURACY
ERM	$87.56 \pm 2.52$	$43.74 \pm 2.11$
MRM	$94.01 \pm 0.82$	<b><math>54.85 \pm 2.11</math></b>
IRM	$88.68 \pm 2.11$	$45.4 \pm 2.40$
MODIRM	$93.01 \pm 0.36$	<b><math>52.35 \pm 1.28</math></b>
REX	$89.85 \pm 1.50$	$47.20 \pm 3.43$
MODREX	$93.55 \pm 1.45$	<b><math>55.51 \pm 2.76</math></b>
DRO	$91.73 \pm 0.40$	$51.95 \pm 1.62$
MODDRO	$92.67 \pm 0.92$	<b><math>55.20 \pm 1.40</math></b>
UNBIAS	$95.00 \pm 0.70$	$72.37 \pm 2.53$



(a) COLOREDOBJECT



(b) SCENEOBJECT

METHODS	TRAIN ACCURACY	TEST ACCURACY
ERM	$98.87 \pm 0.23$	$37.29 \pm 2.74$
MRM	$99.61 \pm 0.04$	<b><math>39.44 \pm 0.77</math></b>
IRM	$98.68 \pm 0.27$	$37.19 \pm 2.58$
MODIRM	$99.39 \pm 0.01$	<b><math>39.14 \pm 1.34</math></b>
REX	$92.91 \pm 1.11$	$38.84 \pm 1.39$
MODREX	$96.71 \pm 0.53$	<b><math>41.04 \pm 1.46</math></b>
DRO	$98.89 \pm 0.35$	$36.34 \pm 1.67$
MODDRO	$99.41 \pm 0.13$	<b><math>39.14 \pm 1.60</math></b>
UNBIAS	$95.25 \pm 2.21$	$56.46 \pm 0.75$



Thank you for listening!