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

Dinghuai Zhang[1] , Kartik Ahuja[1], Yilun Xu[2], Yisen Wang[3], Aaron Courville[1]

[1]MILA, [2]MIT, [3]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

A spurious relationship or spurious correlation is a mathematical relationship in which two or more events or variables are associated but not causally related, due to either coincidence or the presence of a certain third, unseen factor.

## 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 $e \in \mathcal{E} = \{1, \ldots, E\}$

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 $\min_{\theta \in \Theta} \max_{e \in \mathcal{E}} \mathcal{R}^e(\theta)$

## Data structure

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

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

We follow the "realizable" assumption [5], where $Y_e = F(Z^e_{\text{inv}})$

We also assume there exists inverse maps

$$Z^e_{\text{inv}} = G^\dagger_{\text{inv}}(X^e) \quad Z^e_{\text{sp}} = G^\dagger_{\text{sp}}(X^e)$$

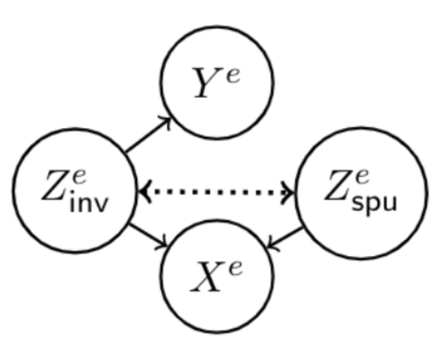
The goal is then to learn

$$F \circ G^\dagger_{\text{inv}}$$

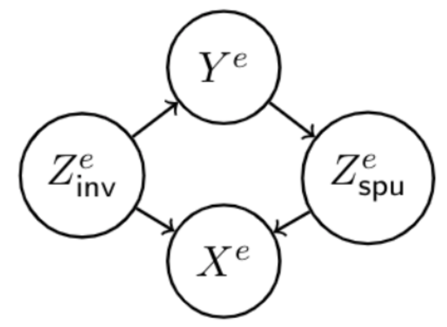
## Related works

"Realizability" means invariant features contain all information about label

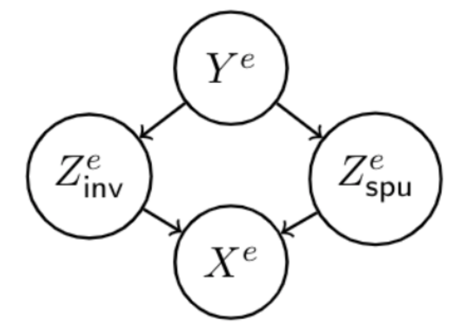
In this work we consider "realizable" case:




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 $Z^e_{\text{sp}}$ and $Y^e$ have a $p^e$ correlation

Side note: When $Z^e_{\text{sp}}$ is high dimensional, the model tends to rely on it

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

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

## Insights

- Sparsity on proper places has good inductive bias for OOD generalization
- In [6], 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 $f(\mathbf{w}_1, \cdots, \mathbf{w}_L; \cdot)$ is parametrized by $\theta = \{\mathbf{w}_1, \cdots, \mathbf{w}_L\}$

We search for a module / subnetwork $f(\mathbf{m}_1 \odot \mathbf{w}_1, \cdots, \mathbf{m}_L \odot \mathbf{w}_L; \cdot)$ 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

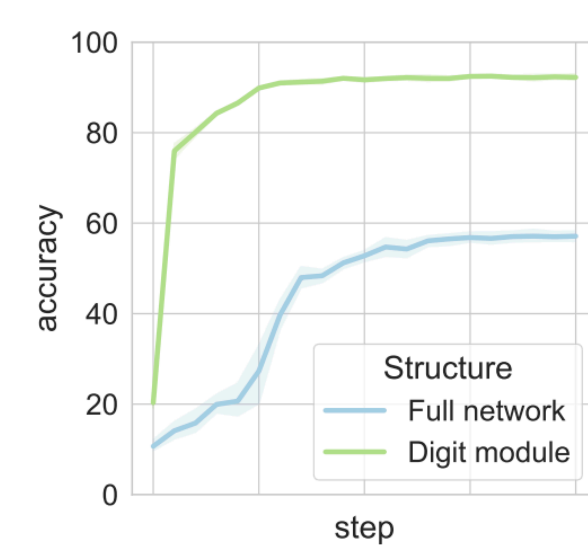
### Modular subnetwork introspection

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:



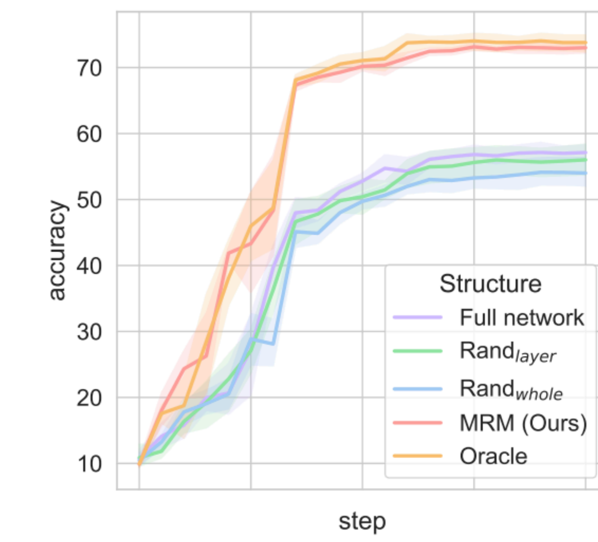

## Functional "lottery ticket"

[4] 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:

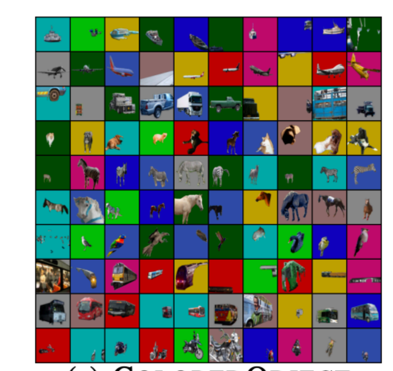
1. train the full model
2. searching module with some desired OOD & sparsity properties
3. retrain the module with same initialization

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



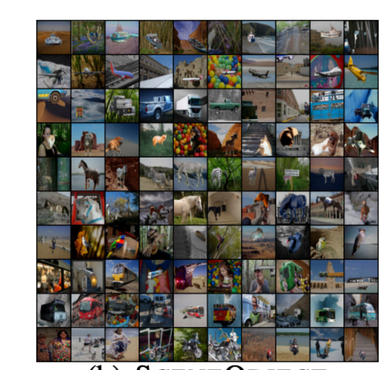

(Oracle means searching module with extra information about test domain in step 2)

## More experiments



(a) ColoredObject

| Methods | Train Accuracy | Test Accuracy |
|---|---|---|
| ERM | 87.56 ± 2.52 | 43.74 ± 2.11 |
| MRM | 94.01 ± 0.82 | **54.85** ± 2.11 |
| IRM | 88.68 ± 2.11 | 45.4 ± 2.40 |
| ModIRM | 93.01 ± 0.36 | **52.35** ± 1.28 |
| REx | 89.85 ± 1.50 | 47.20 ± 3.43 |
| ModREx | 93.55 ± 1.45 | **55.51** ± 2.76 |
| DRO | 91.73 ± 0.40 | 51.95 ± 1.62 |
| ModDRO | 92.67 ± 0.92 | **55.20** ± 1.40 |
| Unbias | 95.00 ± 0.70 | 72.37 ± 2.53 |



(b) SceneObject

| Methods | Train Accuracy | Test Accuracy |
|---|---|---|
| ERM | 98.87 ± 0.23 | 37.29 ± 2.74 |
| MRM | 99.61 ± 0.04 | **39.44** ± 0.77 |
| IRM | 98.68 ± 0.27 | 37.19 ± 2.58 |
| ModIRM | 99.39 ± 0.01 | **39.14** ± 1.34 |
| REx | 92.91 ± 1.11 | 38.84 ± 1.39 |
| ModREx | 96.71 ± 0.53 | **41.04** ± 1.46 |
| DRO | 98.89 ± 0.35 | 36.34 ± 1.67 |
| ModDRO | 99.41 ± 0.13 | **39.14** ± 1.60 |
| Unbias | 95.25 ± 2.21 | 56.46 ± 0.75 |

## References

[1] Invariant Risk Minimization, arxiv 2019

[2] Out-of-Distribution Generalization via Risk Extrapolation, ICML2021

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

[4] The lottery ticket hypothesis: Finding sparse, trainable neural networks. ICLR2019

[5] Out of Distribution Generalization in Machine Learning, arxiv2020

[6] Causal inference using invariant prediction: identification and confidence intervals. Jonas Peters et al. JRSSB