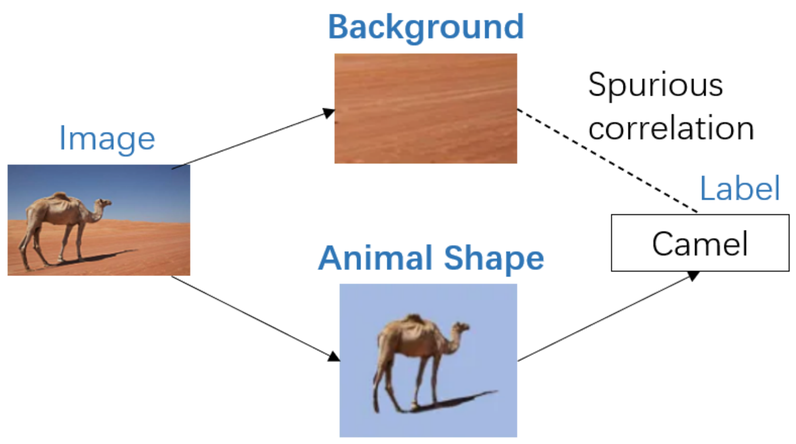
Learning Disentangled Representations for Environment Inference in Out-of-distribution Generalization

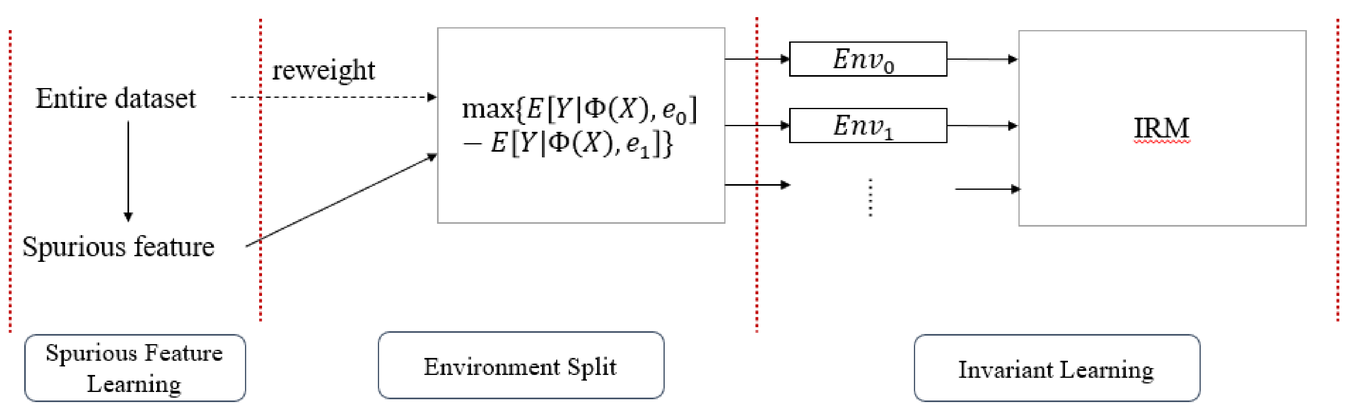
Beijing Jiaotong University, AFCTech, Lenovo

Dongqi Li, Zhu Teng, Qirui Li, Ziyin Wang, Baopeng Zhang, Jianping Fan

Problem

- **Out-of-distribution (OoD) generalization**: Machine learning models may suffer from a sharp drop under a distributional shift.
- Invariant Risk Minimization (IRM): Learn a stable correlation across multiple training environments.
- **Challenges:**
  - **Extra Environment Labels**: IRM requires predefined environment labels, which are not easily accessible.
  - **Improper Reference Model:** spurious features are employed to split datasets, but the quality of spurious features captured by the reference model is insufficient.
- **Our Contributions:**
  - Verifying that ERM-based methods cannot acquire sufficient spurious features.
  - VAE-based method as a reference model to learn disentangled spurious representations.





Proposed Method

- **Observations**

spurious features captured by ERM are insufficient by the two metric:

- Spurious Feature Score
- Invariant Penalty

Reference model	IP↑	SFS↑	Acc(%)↑
ERM	0.0007	0.32	18.8
Ours	**0.1052**	0.75	**63.8**
w/ EnvLabels	0.0025	**1.00**	57.3

- **Learning Disentangled Representations**

The classifier head only takes spurious features as input and the VAE decoder takes both spurious features and residual features as input.

$$ELBO(\phi,\theta,x) = \mathbb{E}_{z\sim q_\phi}\left[log\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \qquad \mathcal{L}(\phi,\theta,\varphi) = -\sum_{(x,y)\in\mathcal{D}_{tr}} ELBO(\phi,\theta,x) + \lambda\cdot log\, p_{\phi,\varphi}(y|x)$$

Experiments

Comparison with Existing Methods

Method	Env Labels	Train Accs	Test Accs
ERM	✘	**89.5 ± 1.1**	26.7 ± 2.8
EIIL	✘	75.2 ± 1.1	59.3 ± 5.5
Ours	✘	81.4 ± 0.3	**69.7 ± 1.8**
IRM	✔	76.7 ± 0.9	70.5 ± 2.2

Table 2: The performance evaluated on Colored MNIST[I].

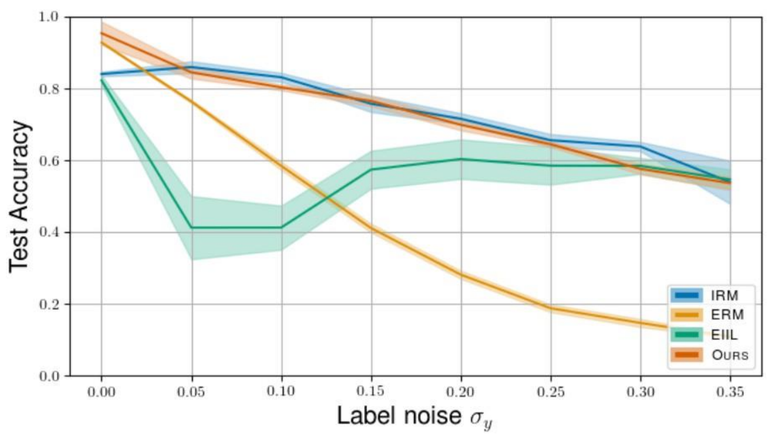
Method	Test Accs
ERM	56.9
EIIL	45.2
Ours	**63.8**
IRM(Oracle)	72.7

Table 3: The performance evaluated on Colored MNIST[II].

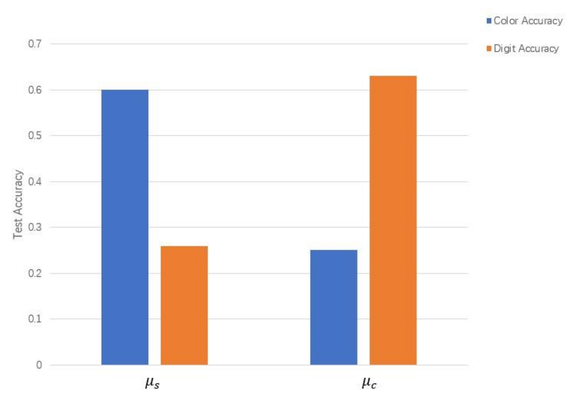
Method	Train Accs	Test Accs
ERM	**93.1 ± 5.9**	71.4 ± 3.4
DisEnt	15.5 ± 1.6	37.4 ± 2.4
ZIN	90.6 ± 0.3	70.8 ± 1.6
EIIL	78.2 ± 12.1	61.9 ± 4.9
Ours	80.4 ± 2.6	**74.4 ± 1.8**
IRM(Oracle)	83.2 ± 1.1	78.7 ± 0.8

Table 4: The performance evaluated on CelebA.

Robustness on label noise



Disentangled Representations Analysis



Acknowledgments

This work was supported by the Natural Science Foundation of China (61972027), the Fundamental Research Funds for the Central Universities of China (2022JBMC009), and the Beijing Municipal Natural Science Foundation (Grant No. 4212041).

BMVC 2023

Paper ID: 538