ICDM 2021
IEEE International Conference on Data Mining
7 – 10 DECEMBER 2021
AUCKLAND NEW ZEALAND

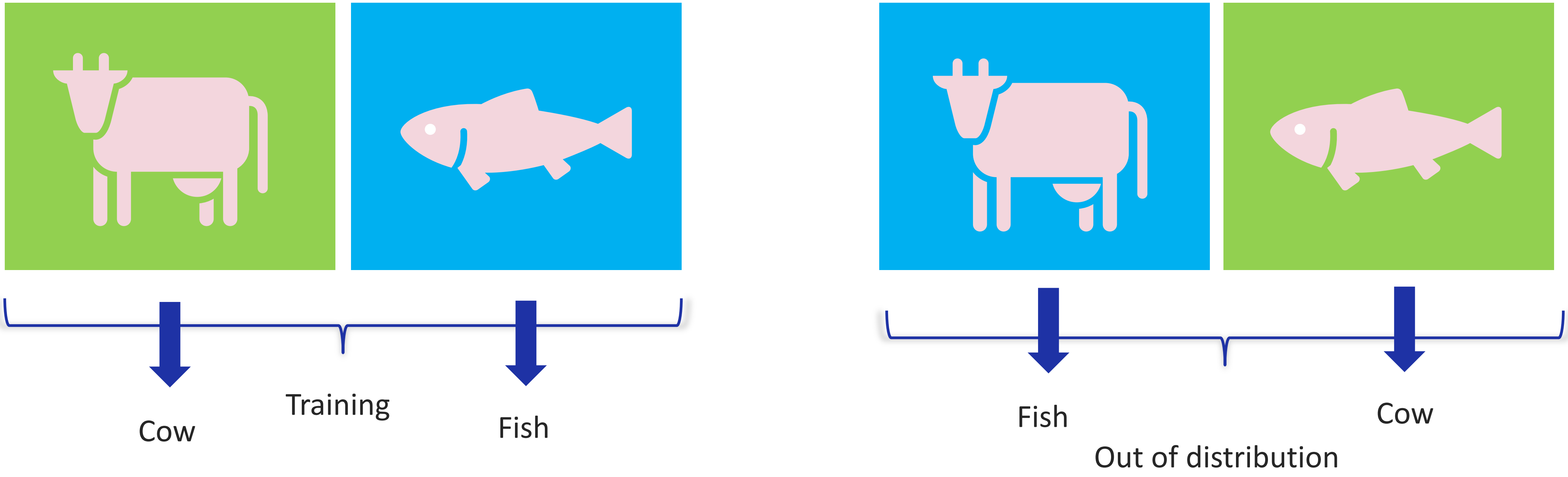
\Orchestrating a brighter world NEC

GIB: Gated Information Bottleneck

for Generalization in Sequential Environments

Francesco Alesiani, Shujian Yu*, Xi Yu**
Machine Learning Group
Data Science and System Platform Division
NEC Laboratories Europe
*UiT - The Arctic University of Norway, *Xi'an Jiaotong University, **University of Florida

Poor Out of Distribution Generalization



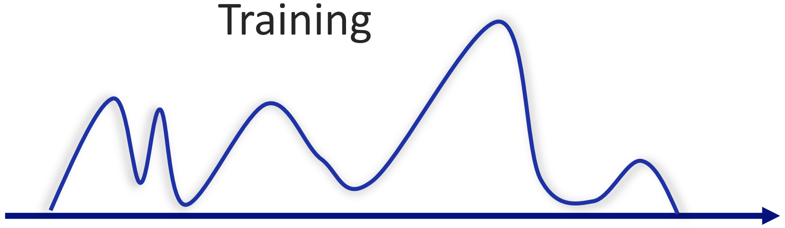
- Deep neural networks also exploit spurious features (e.g. background color)
- Why?

Single Environment

- Independent and Identical distributed (i.i.d.) hypothesis
  - Empirical Risk Minimization (ERM)
- Fails with Out of Distribution generalization

$$\min_{\theta} E_{\{y,x\}} \|y - f_{\theta}(x)\|^2$$

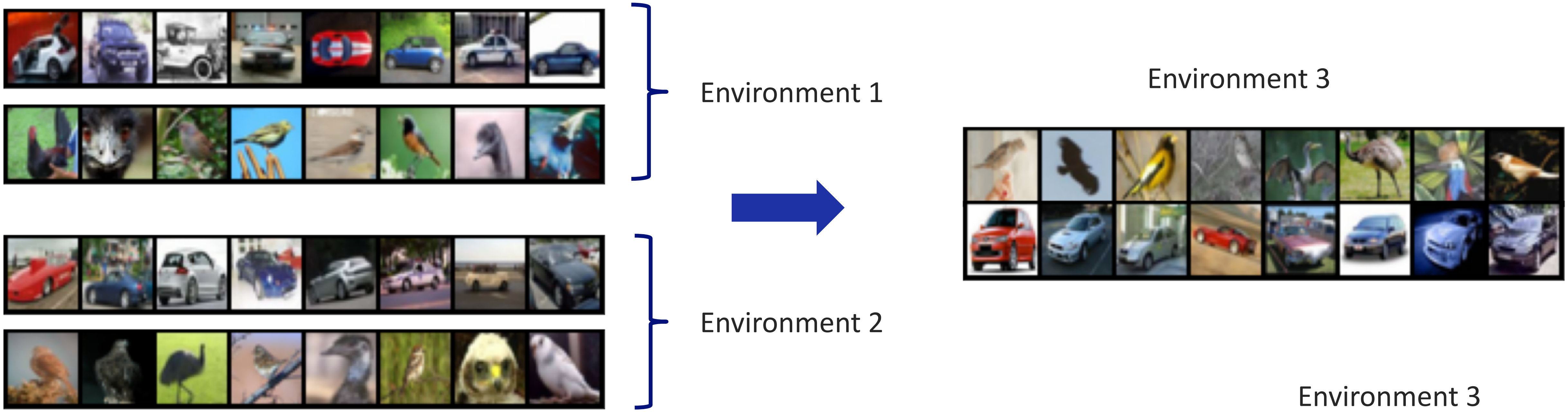
Training

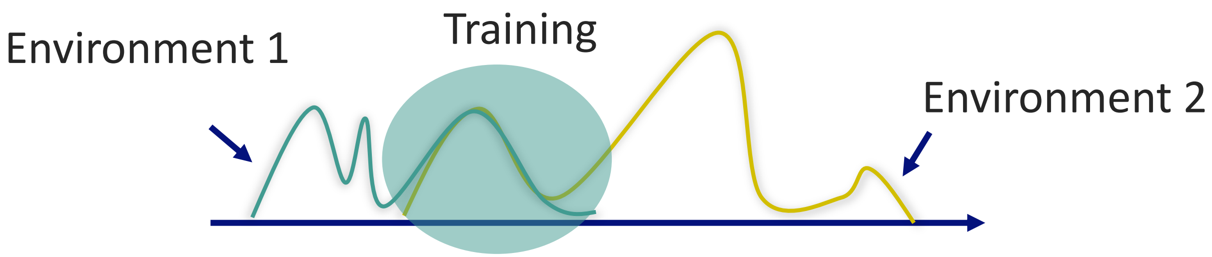


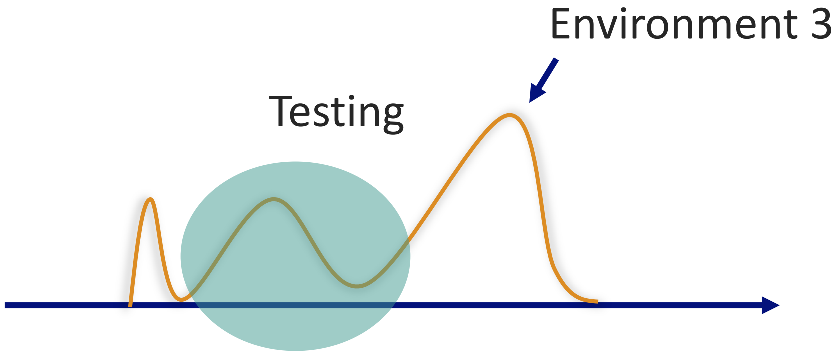
Testing

Multi Environments

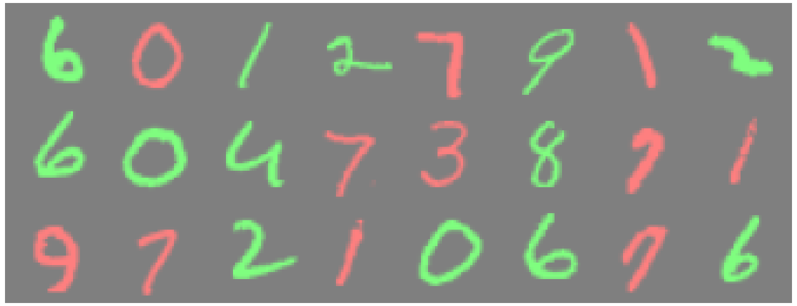
- Causal model hypothesis for Out-of-Distribution generalization
  - E.g. Invariant Risk Minimization (IRM) [Arjovsky et al. 2019]

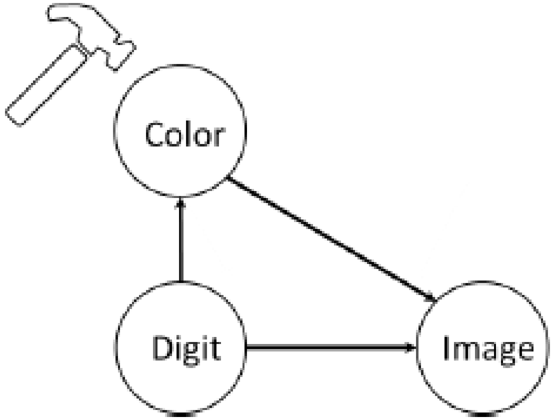






MNIST digit example





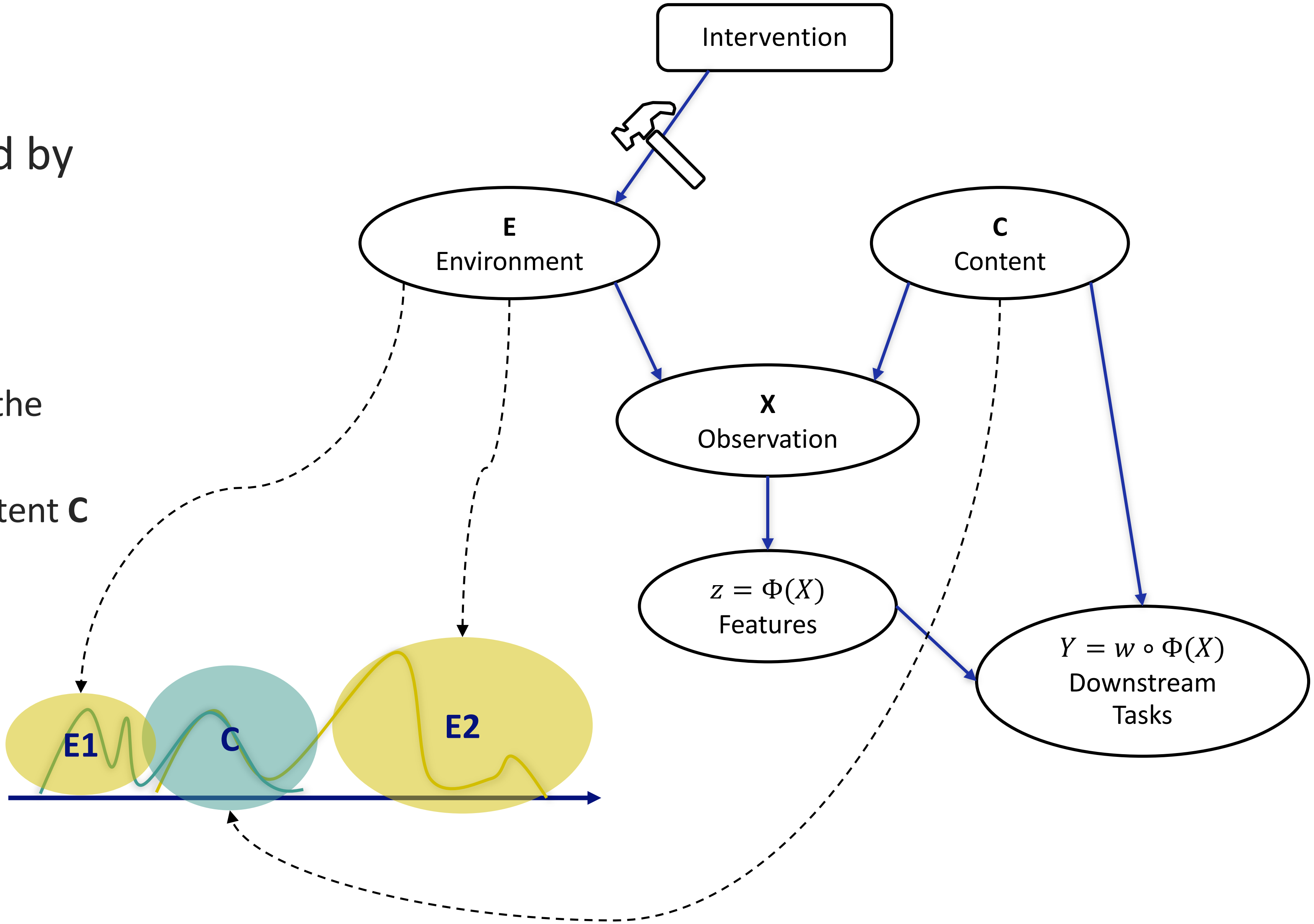
- Stable feature: digit shape
- Spurious feature: digit color, depends on the digit value

  NEC Laboratories Europe

\Orchestrating a brighter world

Multi Environments

- Causal Model
- Observation are generated by
  - Content (**C**)
  - Environment (**E**)
- Feature extraction $\boldsymbol{\Phi}$
  - remove the contribution from the environment **E**
  - provide information of the content **C**



  NEC Laboratories Europe \Orchestrating a brighter world NEC

Information Bottleneck (IB)

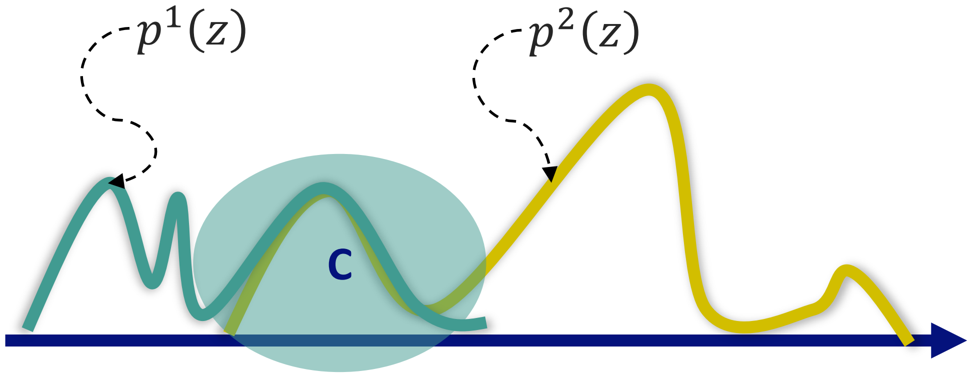
$$X \rightarrow Z \rightarrow Y \qquad\qquad L_{IB} = I(Y;Z) - \lambda I(X;Z)$$

- Information Theory
- Maximize the Mutual Information (MI) of the latent Z and the target variable Y
  - i.e. Z maximal informative
- Minimize Mutual Information between source X and latent Z
  - i.e. remove spurious features
- λ hyper parameter

Principle

Only common features

$$\min_{\phi} \mathrm{KL}(p^1(z) || p^2(z))$$



Using sample distributions

$$\min_{\phi} \mathbb{E}_{x,x'} \mathrm{KL}(p^1(z|x) || p^2(z|x'))$$

- Learn common or invariant features among environments

Gated Information Bottleneck

- Parallel environments

$$\min_{\Phi,w} \sum_{e\in E} R^e(w\circ\Phi) \ + \lambda \sum_{e\in E} \sum_{e'\in E} \mathbb{E}_{x\sim p^e(x)} \mathbb{E}_{x'\sim p^{e'}(x)} \mathrm{KL}(p(z|x)||p(z|x'))$$

Lemma 2 (Cross-Domain Mutual Information Upper Bound).

$$I^{12}(X;Z) \le \mathbb{E}_x \mathbb{E}_{x'} \mathrm{KL}(p^1(z|x)||p^2(z|x'))$$

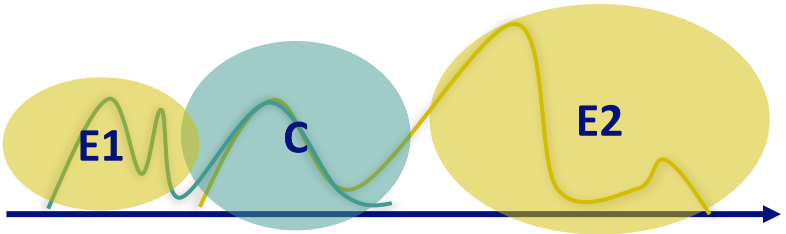
$$\min_{\Phi,w} \sum_{e\in E} R^e(w\circ\Phi) + \lambda I^e(X;Z)$$

For deterministic encoder $X \to Z$

$$I(X;Z) = H(Z) - H(Z|X) = H(Z)$$

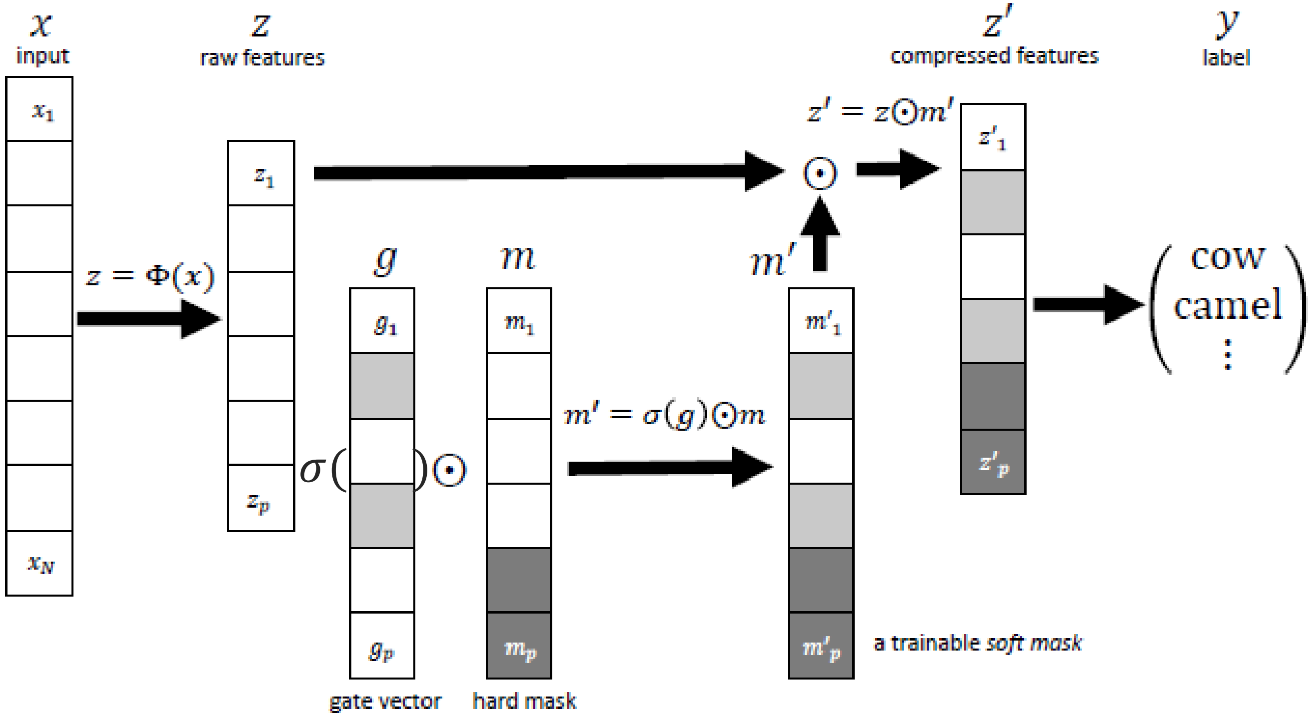
$$\min_{\Phi,w} \sum_{e\in E} R^e(w\circ\Phi) + \lambda H^e(Z)$$

Gated Information Bottleneck



- sequential environments

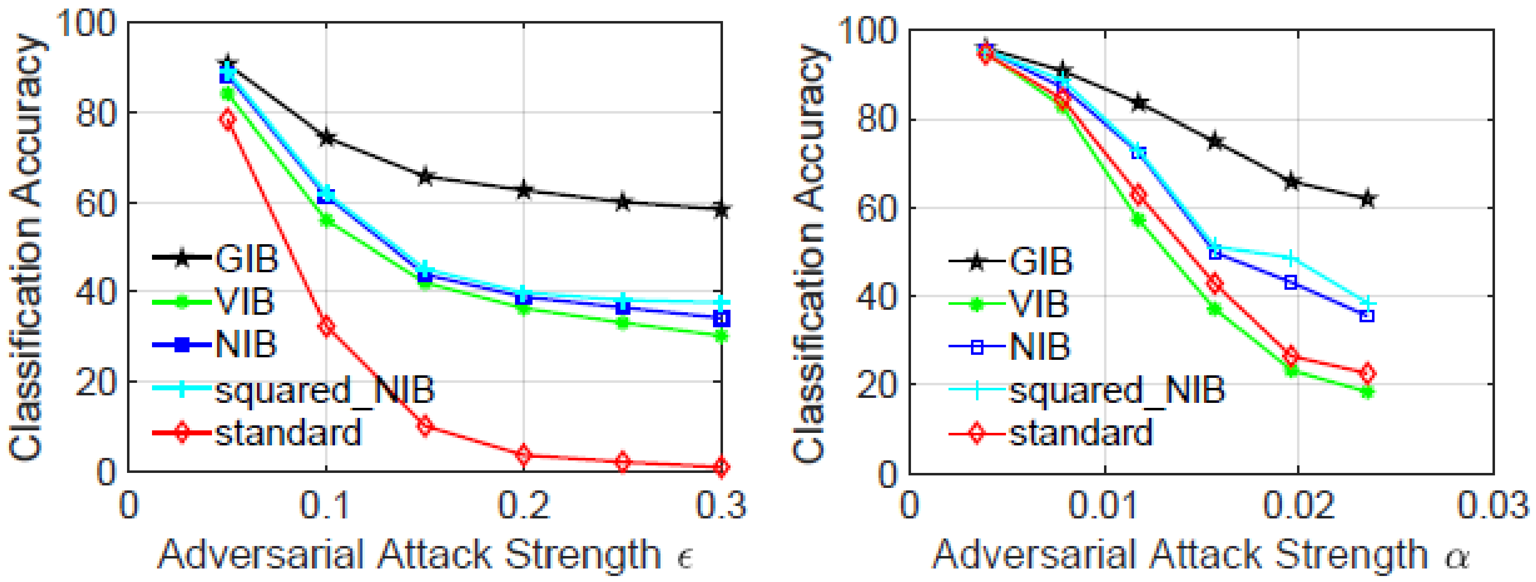
$$\min_{\Phi,w} \sum_{e\in E} R^e(w\circ\Phi) + \lambda H^e(Z)$$



$$m' = m\odot\sigma(g) \qquad \Phi' = m'\odot\Phi \qquad z = \Phi'(x) \qquad m = 1_{\{\sigma(g)\geq\tau\}} \text{ or } m = 1_{\{\sigma(g)\geq\tau\wedge m=1\}}$$

Experiments – Parallel Environments

- Robustness to adversarial attacks



(a) FGSM (b) PGD

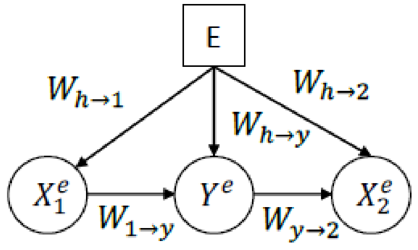
- Out-of-Distribution Detection (OoDD)

Methods	Standard	VIB	NIB	squared NIB	GIB
AUROC↑	90.2	90.6	91.6	92.1	**93.0**
AUPR In↑	93.3	92.1	93.1	93.1	**94.5**
AUPR Out↑	90.5	90.3	91.3	**91.5**	91.3
Detection Acc↑	83.2	83.2	84.1	84.1	**86.9**
FPR (95% TPR)↓	49.6	48.0	49.3	49.4	**47.9**

Experiments – Sequential Environments

- Synthetic dataset
- Colored MNIST, Fashion-MNIST, KMNIST, EMNIST
- Multiple environments

Method	Causal Error	Non Causal Error
SEM (ground true)	0.00 ± 0.00	0.00 ± 0.00
SERM	92.0 ± 1.9	92.9 ± 1.6
SIRM	66.2 ± 1.6	65.5 ± 1.0
GIB	**13.7** ± 0.4	**42.5** ± 0.7

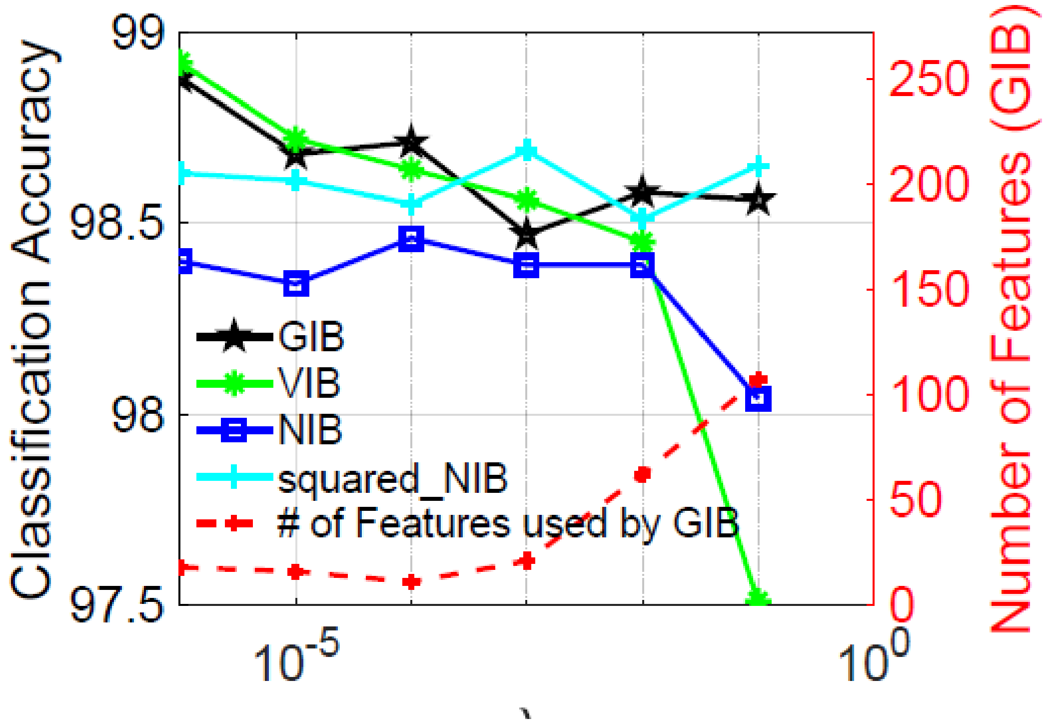


Dataset	FaMNIST		KMNIST		EMNIST	
Method	train acc	test acc	train acc	test acc	train acc	test acc
EWC	82.7% (0.5%)	24.4% (1.1%)	82.7% (0.4%)	21.1% (1.3%)	82.7% (0.2%)	21.1% (0.6%)
GEM	82.4% (0.3%)	24.5% (1.6%)	83.1% (0.4%)	20.8% (1.5%)	82.9% (0.6%)	21.3% (0.6%)
MER	78.7% (0.4%)	19.6% (2.1%)	80.9% (0.5%)	20.1% (1.7%)	78.8% (1.0%)	19.3% (1.9%)
ERM	82.7% (0.5%)	24.4% (1.2%)	82.7% (0.4%)	21.1% (1.3%)	82.7% (0.2%)	21.1% (0.6%)
IRM	82.7% (0.5%)	24.1% (1.0%)	82.7% (0.4%)	21.1% (1.3%)	82.7% (0.2%)	21.1% (0.6%)
IRMG	84.0% (0.8%)	26.4% (1.4%)	84.3% (0.1%)	23.9% (1.2%)	83.8% (0.8%)	23.9% (0.6%)
GIB	79.9% (5.1%)	**55.1%** (3.8%)	65.3% (12.4%)	**47.5%** (3.7%)	66.2% (1.4%)	**49.3%** (2.4%)

Number Env.	2		4		6	
Method	train acc	test acc	train acc	test acc	train acc	test acc
EWC	83.0% (0.6%)	24.4% (1.3%)	80.0% (0.4%)	24.7% (0.9%)	78.9% (0.6%)	22.9% (0.3%)
GEM	83.0% (0.4%)	24.9% (1.2%)	80.2% (0.5%)	24.6% (1.3%)	79.1% (0.6%)	23.8% (0.5%)
MER	78.0% (1.0%)	24.4% (3.0%)	77.2% (0.6%)	22.8% (2.4%)	76.5% (0.6%)	24.5% (3.4%)
ERM	83.0% (0.6%)	24.4% (1.3%)	80.1% (0.4%)	24.7% (0.8%)	78.9% (0.6%)	22.9% (0.3%)
IRM	83.1% (0.6%)	24.8% (1.3%)	74.8% (0.6%)	16.5% (4.3%)	74.7% (0.7%)	13.9% (4.6%)
IRMG	83.8% (0.6%)	26.7% (0.6%)	79.4% (0.2%)	28.1% (0.9%)	77.2% (0.4%)	27.9% (0.5%)
GIB	75.7% (2.6%)	**55.1%** (2.1%)	66.1% (11.3%)	**52.9%** (2.9%)	69.0% (1.8%)	**53.3%** (1.8%)

Feature analysis

- How many features are relevant?



Related works

- Continual Invariant Risk Minimization [Alesiani et al. 2020, ICLR Workshop]
  - Motivation of the current work
- Drop-Bottleneck [Kim et al. 2021, ICLR]
  - Probabilistic feature selection
- Representation learning via invariant causal mechanism [Mitrovic et al. 2020, ICLR]
  - Image representation learning using augmentation and invariant learning
- IRM, IRMG, …

Conclusions

- We propose the use of gated features to learn invariant feature extraction
- Gate is deterministic
- Improves both robustness against adversarial attacks and out of distribution detection
- Shows favorable performance in sequential environments
- Connection between Information Bottleneck and Invariant Risk minimization