FLOOD: A Flexible Invariant Learning Framework for Out-of-Distribution Generalization on Graphs

Yang Liu[1]; Xiang Ao[1*]; Fuli Feng[2]; Yunshan Ma[3]; Kuan Li[1]; Tat-Seng Chua[3]; Qing He[1*]

柳 阳[1]； 敖 翔[1*]； 冯福利[2]； 马云山[3]； 李宽[1]； 蔡达成[3]； 何 清[1*]

1 中国科学院计算技术研究所 INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

2 中国科学技术大学 University of Science and Technology of China

3 NUS National University of Singapore

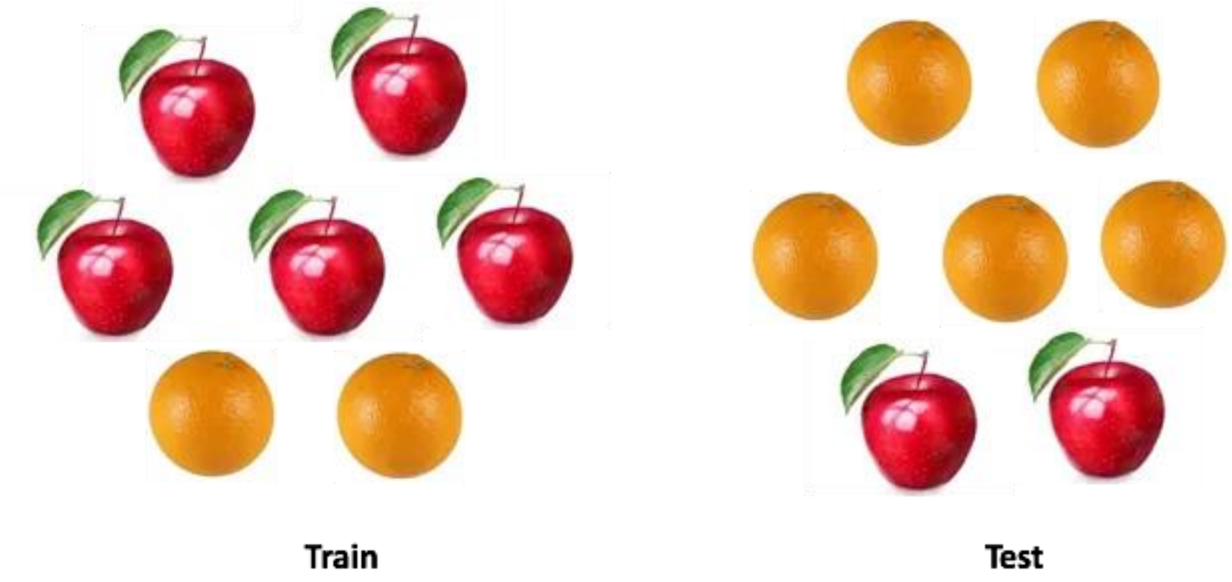
* denotes corresponding author.

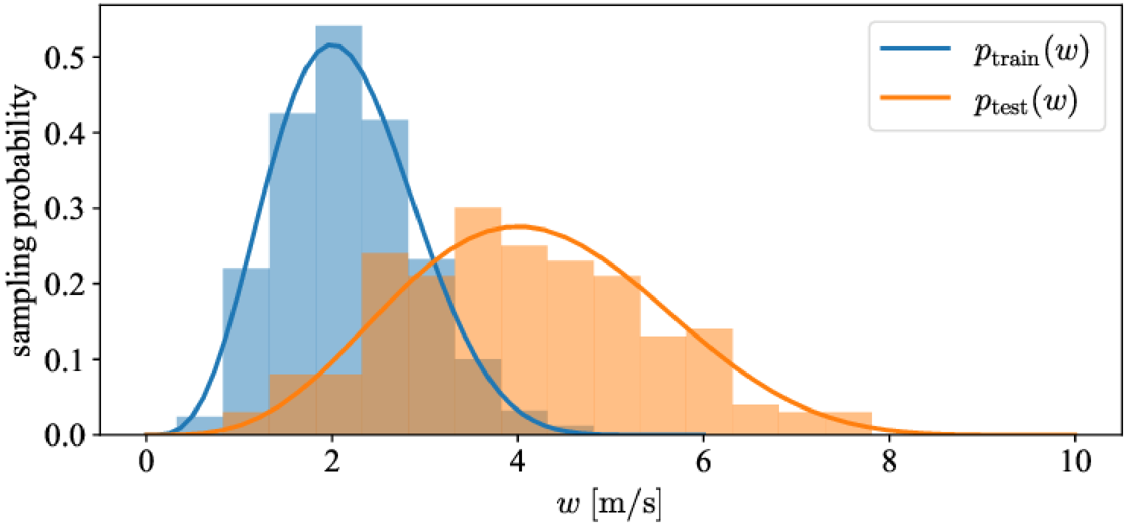
Content

- Background and Motivation
- Method – FLOOD
- Experiment
- Conclusion

Background

- Some simple questions about train and test:
  - In real-world scenarios, how can we know whether the training data and testing data follow the same distribution?
  - If there is a slight difference between training and testing set, can the model still achieve good generalization performance?

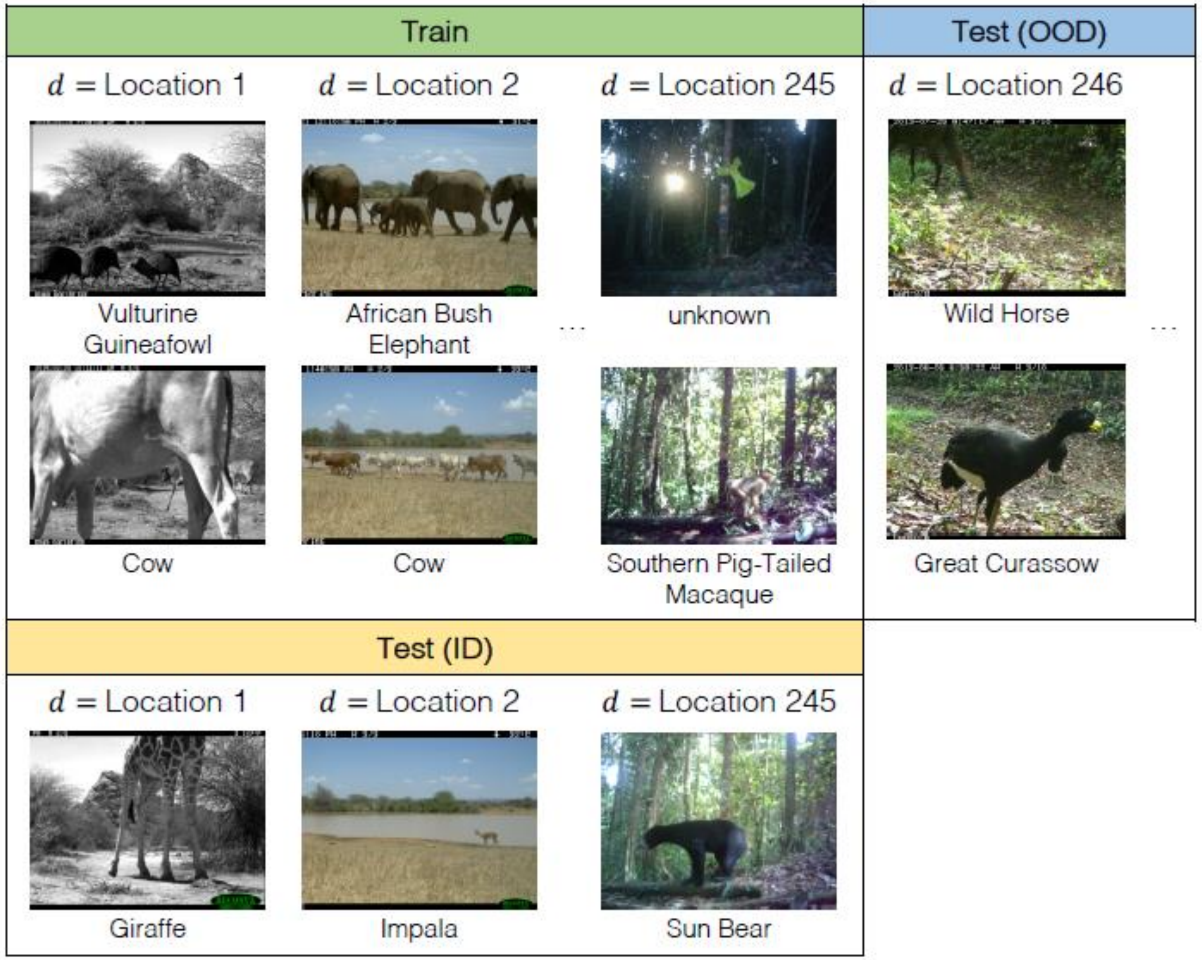


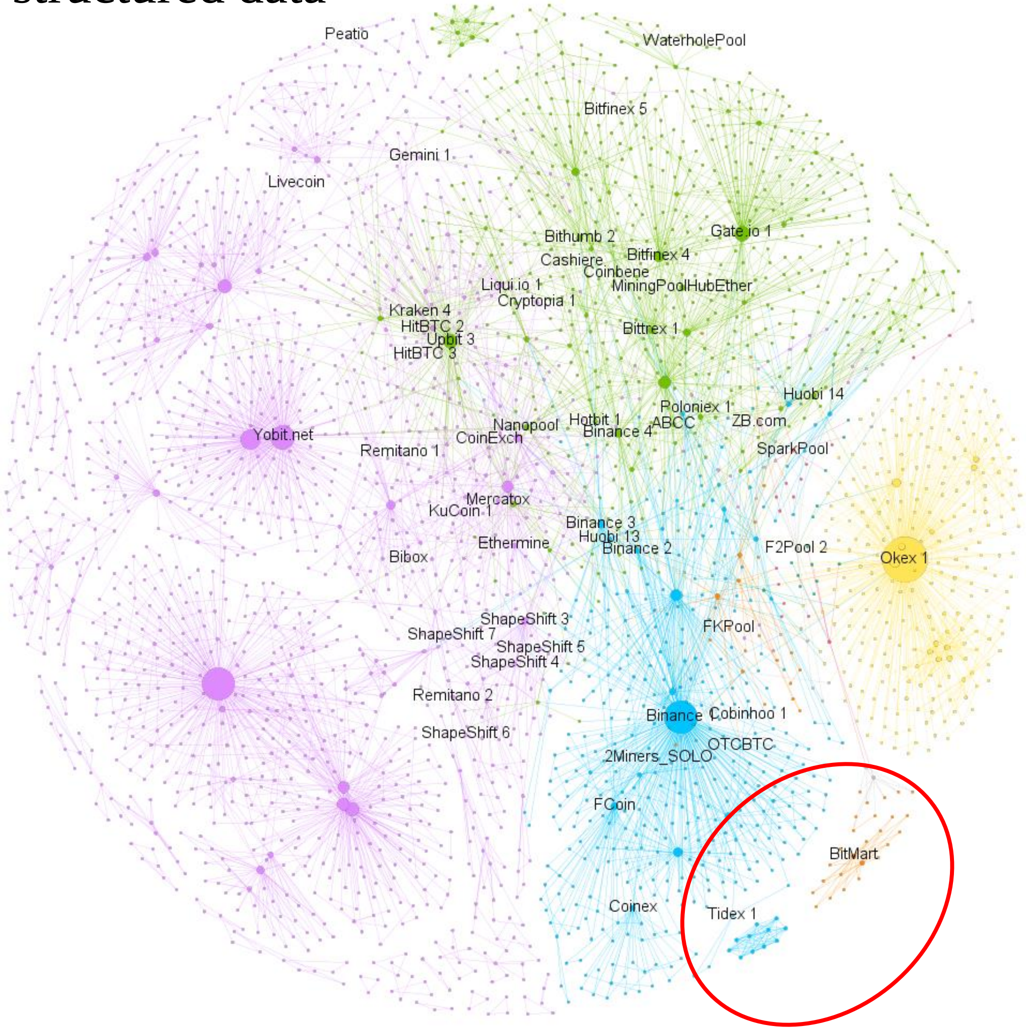


Shikhar Gupta, How (dis)similar are my train and test data? in Towards Data Science

Background

- **Out-of-distribution** data are **ubiquitous** in real-world situations
  - Unlike images, OOD samples are ambiguous for graph-structured data

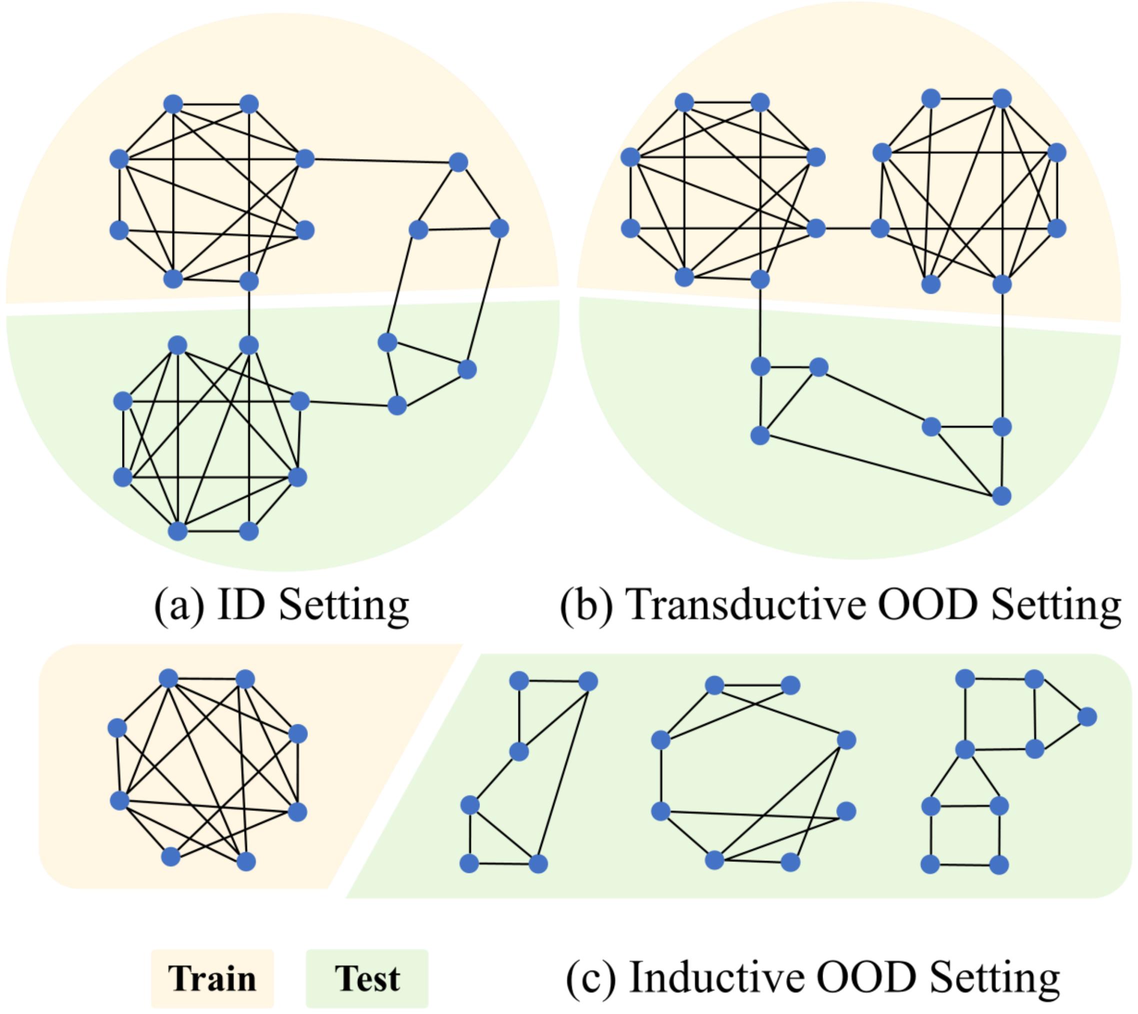




Koh et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts, in ICML'21

Background

- **In-distribution v.s. Out-of-distribution**
  - OOD can be defined in terms of certain node attribute like **node degree.**
  - In-distribution: The training nodes and testing nodes follow **similar** degree distribution.
  - Transductive OOD: The degree of testing nodes is **different** from that of training nodes.
  - Inductive OOD: The training nodes and testing nodes are from different graphs, thus follow **different** distributions.



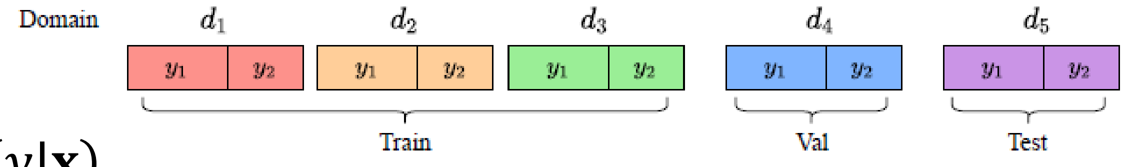
Preliminaries

➢ Distribution shift

- $\mathrm{P}(\mathbf{x}, y) = \mathrm{P}(\mathbf{x})\mathrm{P}(y|\mathbf{x})$

- Covariate Shift

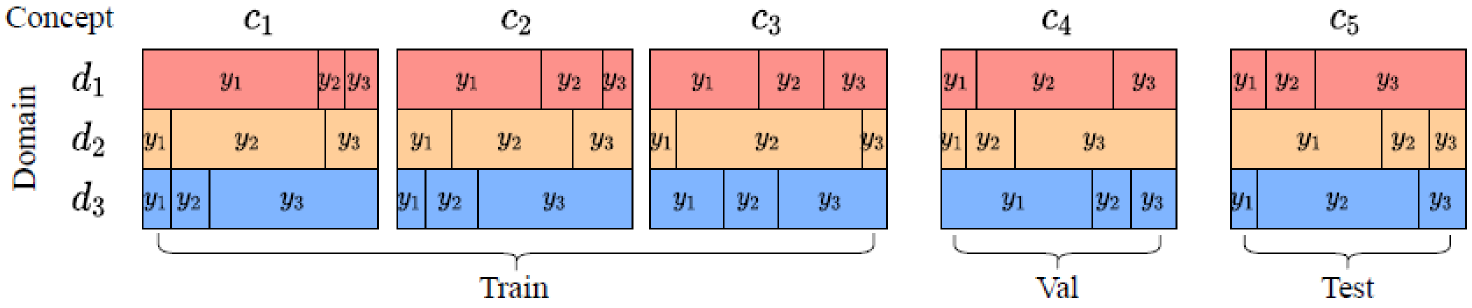
$$P_{\text{train}}(\mathbf{x}) \neq P_{\text{test}}(\mathbf{x}) \text{ and } P_{\text{train}}(y|\mathbf{x}) = P_{\text{test}}(y|\mathbf{x})$$



(a) Covariate shift split

- Concept Shift

$$P_{\text{train}}(\mathbf{x}) = P_{\text{test}}(\mathbf{x}) \text{ and } P_{\text{train}}(y|\mathbf{x}) \neq P_{\text{test}}(y|\mathbf{x})$$



(c) Concept shift split

Gui et al., GOOD: A Graph Out-of-Distribution Benchmark, in NeurIPS'22

Preliminaries

➢ Invariant learning

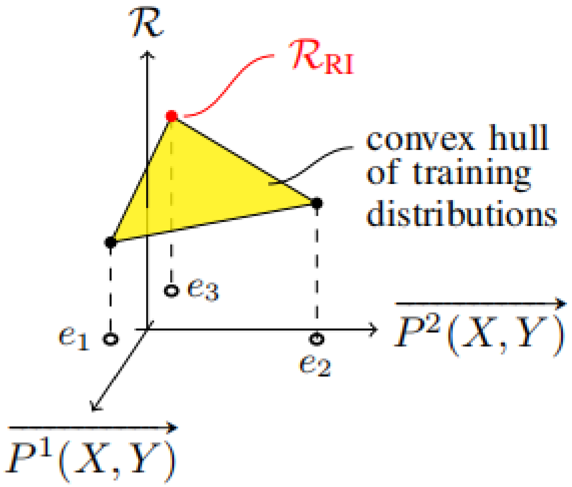
- Invariant Risk Minimization (IRM)

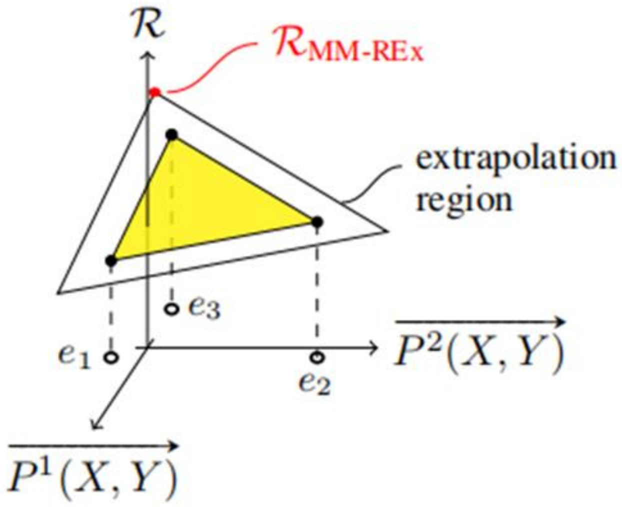
$$\mathcal{R}_{\text{IRM}}(\psi) = \sum_{e \in \mathcal{E}^{obs}} \mathcal{R}_e(\psi) + \lambda \|\nabla_\omega \mathcal{R}_e(\omega \circ \psi)\|$$

- Risk Extrapolation (REx)

$$\mathcal{R}_{\text{REx}}(\psi) = \max_{\substack{\sum_e \lambda_e = 1 \\ \lambda_e \geq \lambda_{\min}}} \sum_{e=1}^{M} \lambda_e \mathcal{R}_e(\psi)$$

$$\mathcal{R}_{\text{e}}(\psi) = \mathcal{R}_{\text{ERM}}(\psi) = \mathbb{E}_{P\,(x,y,e)}[\ell(f_\psi(x), y)]$$

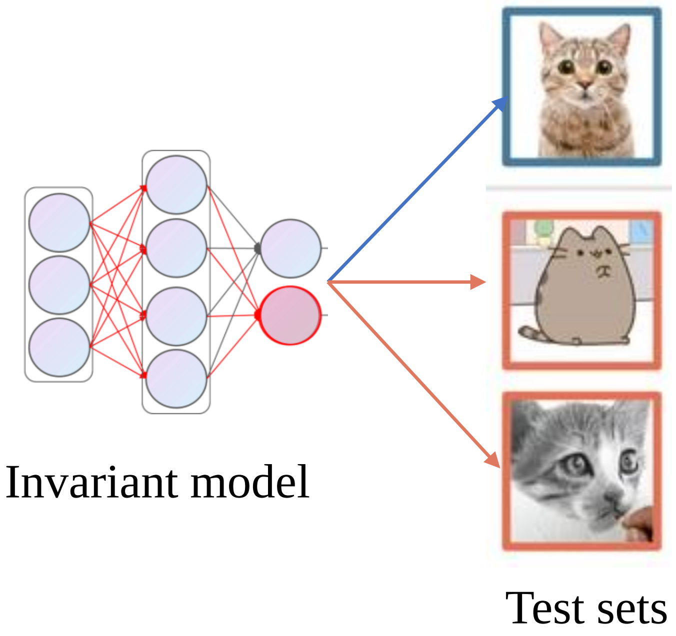


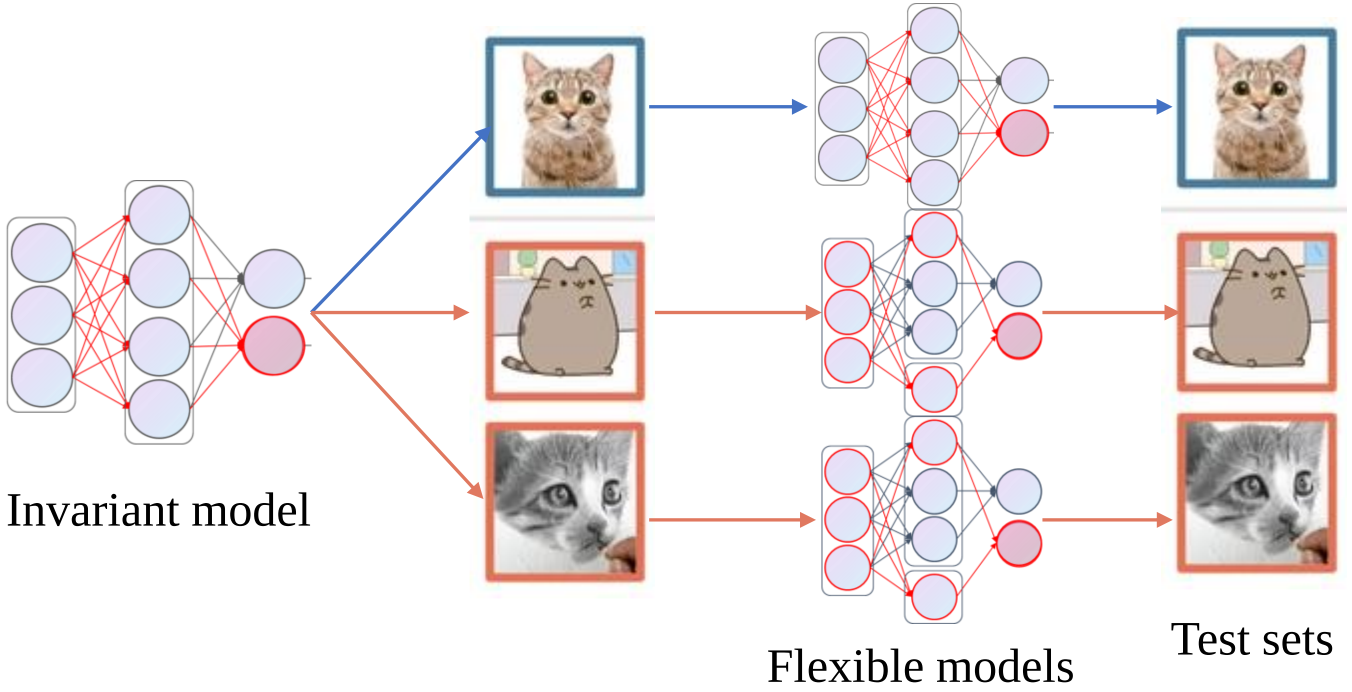


Krueger et al., Out-of-Distribution Generalization via Risk Extrapolation, in ICML'21

Motivation

- **Existing invariant learning approaches are not flexible to tackle the graph OOD problem**
  - Expect one model to generalize to various test distributions
  - Cope with distribution shift by sticking to an invariant principle
  - A better solution is to adapt the model to the target distribution flexibly

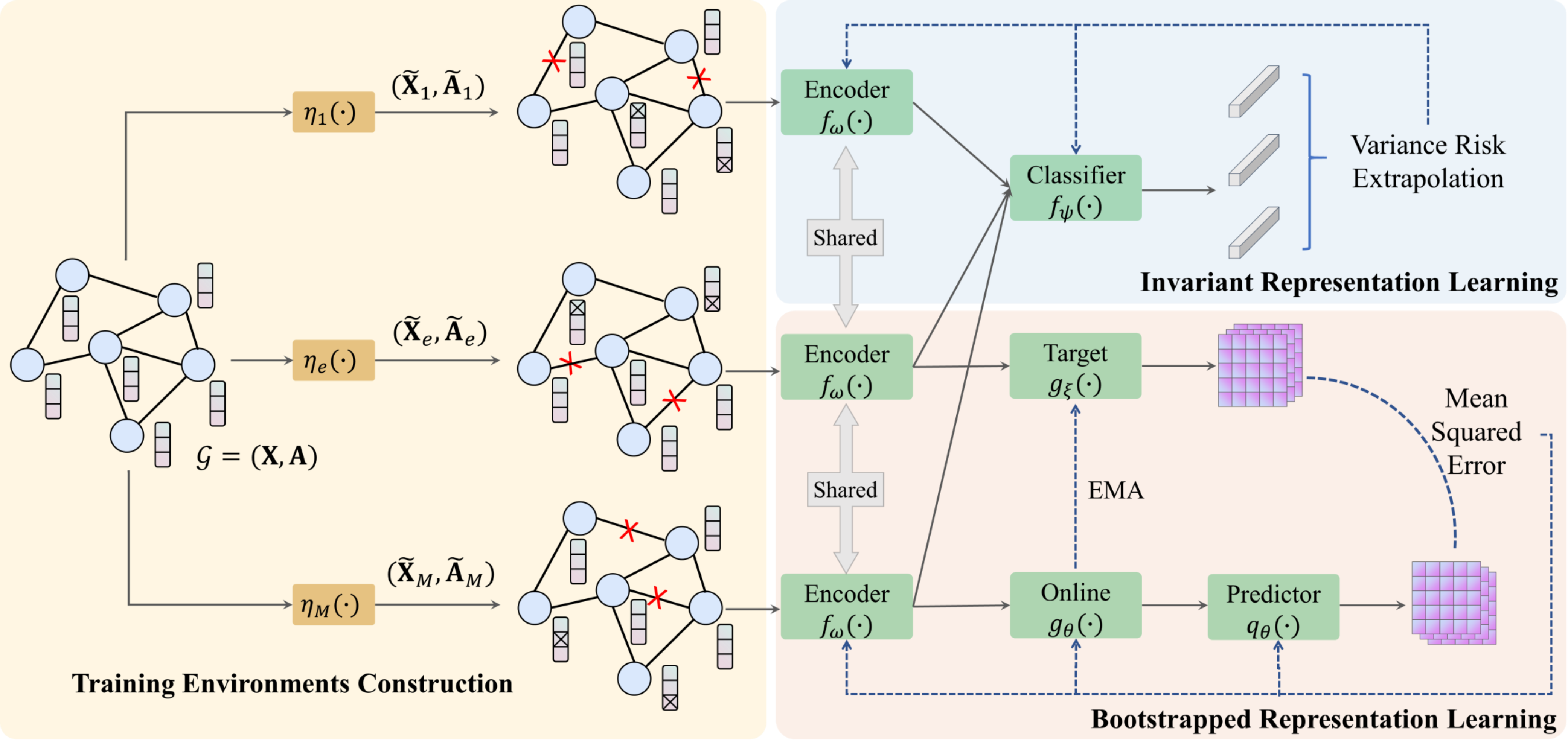




Content

- Background and Motivation
- **Method – FLOOD**
  - Training Environments Construction
  - Invariant Representation Learning
  - Bootstrapped Representation Learning
- Experiment
- Conclusion

➢ **FLOOD:** Flexible invariant Learning framework for Out-Of-Distribution generalization on graphs



➢ Training Environments Construction

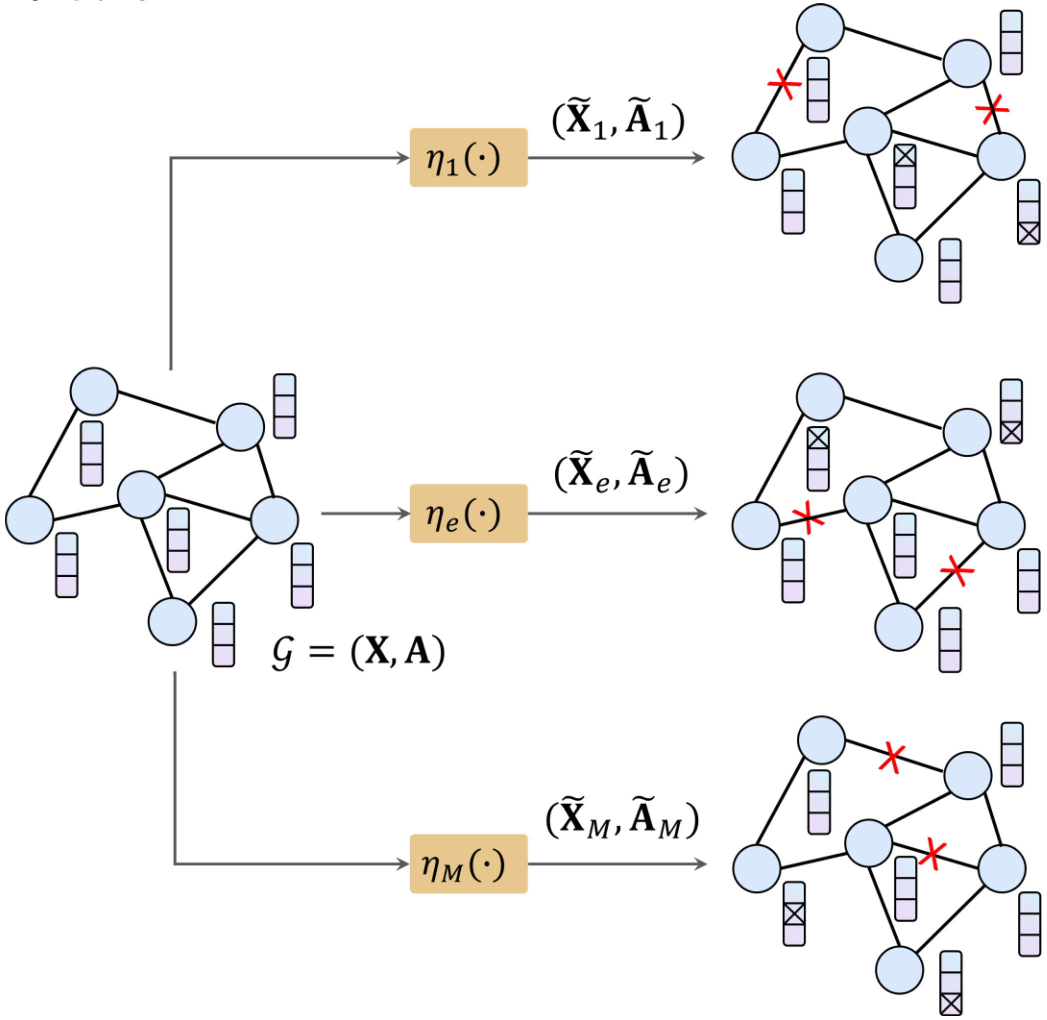
- Node feature masking:

$$o_v^M \in \{0,1\}^d$$

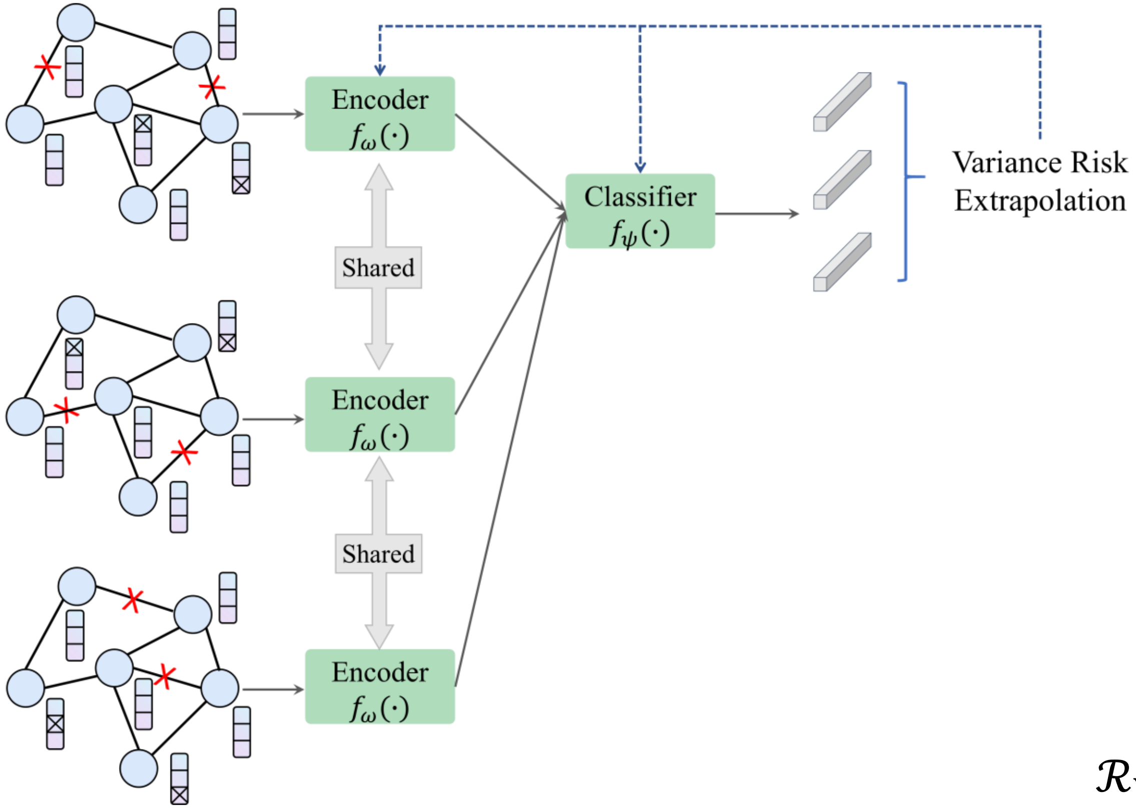
- Edge dropping:

$$o_{(u,v)}^E \in \{0,1\}$$

$$\eta_e(\mathrm{X}, \mathrm{A}) = \left(\widetilde{\mathrm{X}}_e, \widetilde{\mathrm{A}}_e\right), \quad e=1 \ldots M$$



➢ Invariant Representation Learning



- Variance Risk Extrapolation
  - reducing training risks while increasing the similarity of training risks to improve generalization on target distribution

$$\mathcal{R}_e(\omega,\psi) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} Y_{ij}\log\left[f_\psi[f_\omega(\tilde{\mathrm{X}}_e,\tilde{\mathrm{A}}_e)]\right]_{ij}$$

$$\mathcal{R}_{\mathrm{REx}}(\omega,\psi) = \max_{\substack{\sum_e \lambda_e = 1\\ \lambda_e \geq \lambda_{\min}}} \sum_{e=1}^{M} \lambda_e \mathcal{R}_e(\omega,\psi)$$

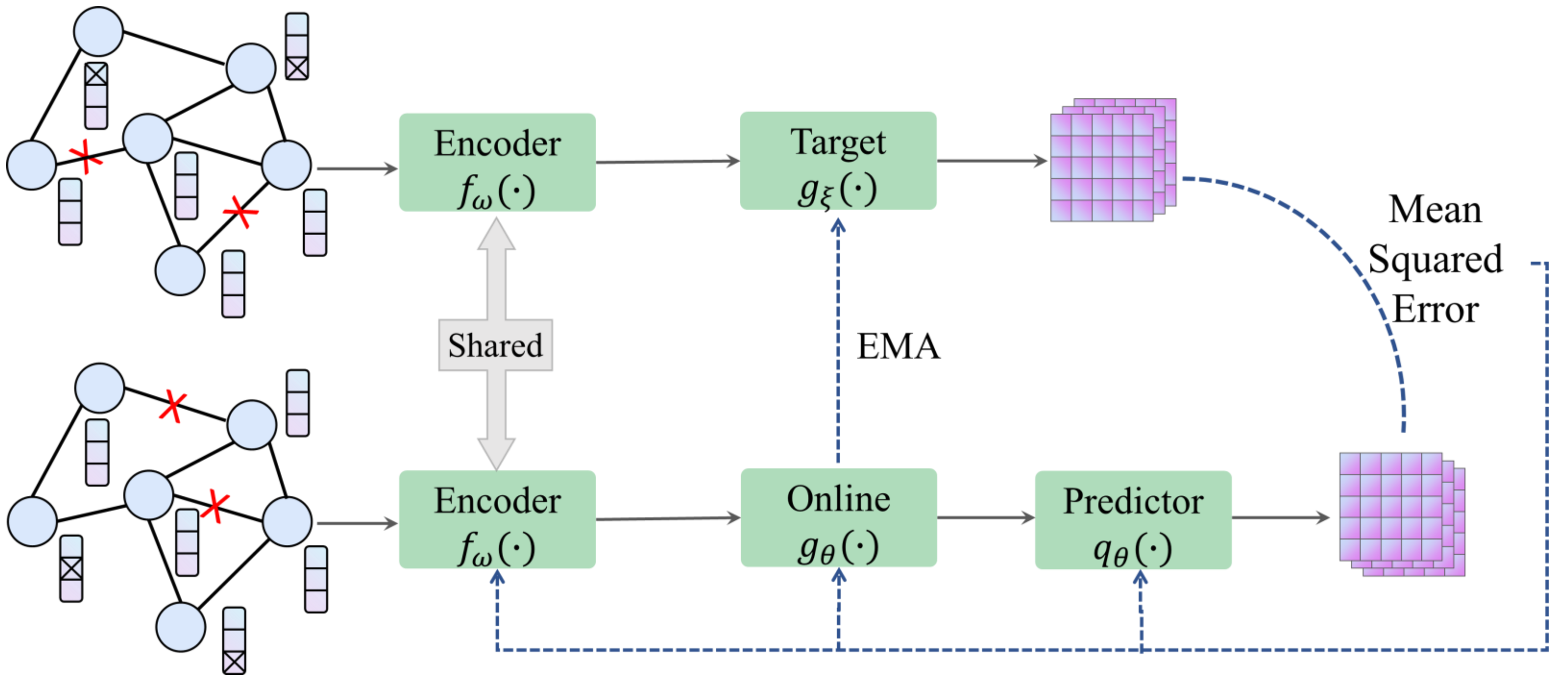
$$\mathcal{R}_{\mathrm{V-REx}}(\omega,\psi) = \beta\cdot\mathrm{Var}([\mathcal{R}_1(\omega,\psi),\ldots,\mathcal{R}_M(\omega,\psi)]) + \sum_{e=1}^{M}\mathcal{R}_e(\omega,\psi)$$

➢ Bootstrapped Representation Learning

- online view: $\mathbf{Z}_\theta = g_\theta\left(f_\omega(\widetilde{\mathrm{X}}_i, \widetilde{\mathrm{A}}_i)\right)$
- target view: $\mathbf{Z}_\xi = g_\xi\left(f_\omega(\widetilde{\mathrm{X}}_j, \widetilde{\mathrm{A}}_j)\right)$

$$\mathcal{L}(\theta, \omega) = \frac{1}{N}\sum_{k=1}^{N}\left\|\frac{q_\theta(\mathbf{Z}_{\theta,k})}{\left\|q_\theta(\mathbf{Z}_{\theta,k})\right\|_2} - \frac{\mathbf{Z}_{\xi,k}}{\left\|\mathbf{Z}_{\xi,k}\right\|_2}\right\|_2^2$$

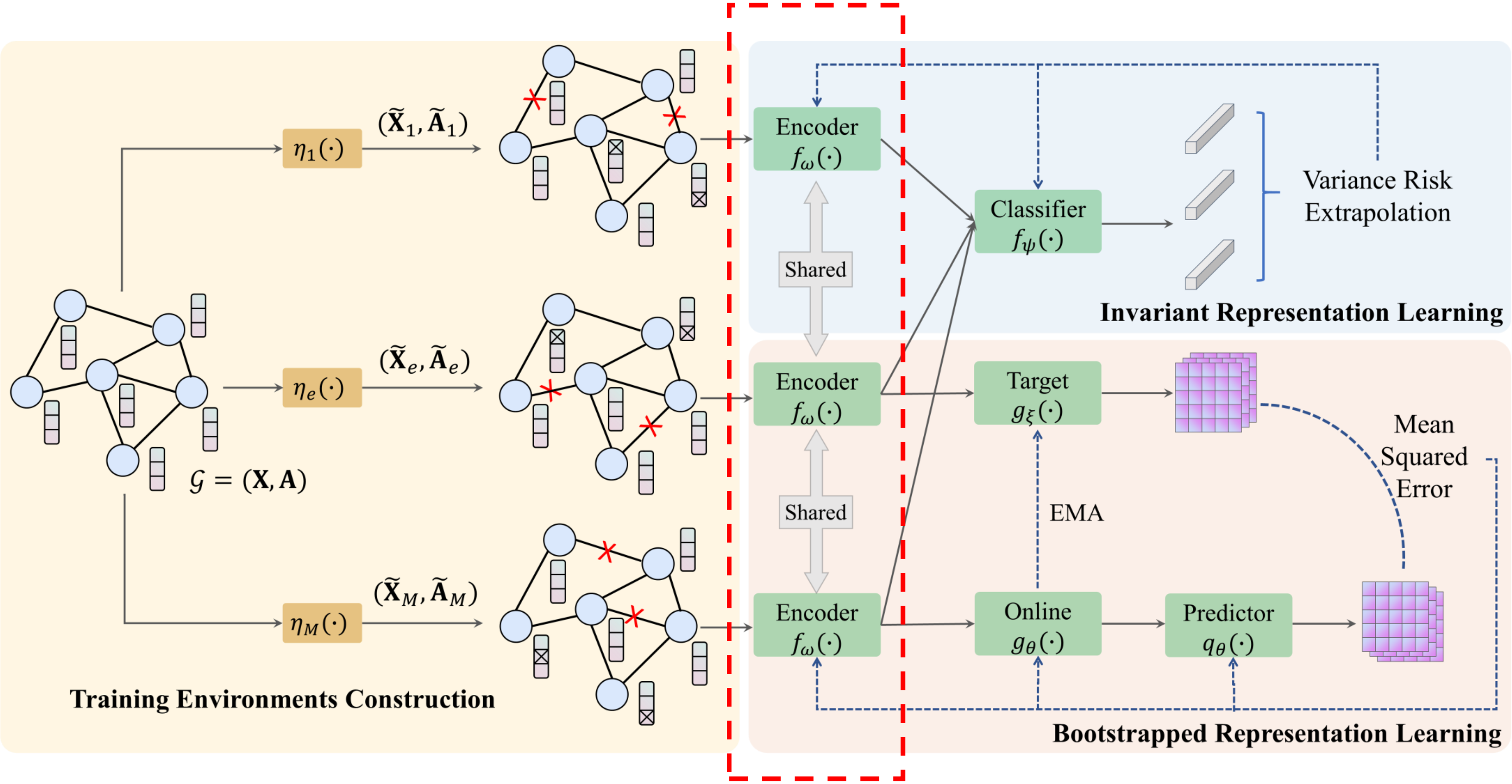
$$\xi \leftarrow \tau\xi + (1-\tau)\theta$$



➢ Test-time training

$$\min_{\theta,\omega,\psi} \mathcal{L}_{\text{train}} = \mathcal{L}(\theta,\omega) + \alpha\mathcal{R}_{\text{V-REx}}(\omega,\psi)$$

$$\min_{\omega} \mathcal{L}_{\text{test}} = \mathcal{L}(\theta,\omega)$$



Content

- Background and Motivation
- Method – FLOOD
- **Experiment**
  - RQ1: Does FLOOD outperform the state-of-the-art methods for out-of-distribution generalizations on graphs?
  - RQ2: How do the key components contribute to the results?
  - RQ3: How does test-time training improve the generalization of GNNs?
  - RQ4: What is the sensitivity of FLOOD with respect to the number of training environments and gradient descent steps during the test phase?
- Conclusion

Data

➤ Public benchmark

- **GOOD-CBAS**: colored BA-Shapes
- **GOOD-WebKB**: a graph of university webpage
- **GOOD-Cora:** a citation network labeled on the paper topic
- **GOOD-Arxiv:** the citation network between all Computer Science (CS) arXiv papers indexed by MAG
- **Twitch-explicit**: contains 6 networks where Twitch users are nodes, and mutual friendships between them are edges

Table 1: Statistics of GOOD datasets for transductive tasks.

Dataset	#Node	#Edge	#Class	#Feat	Domain
CBAS	700	3,962	4	4	Color
WebKB	617	1,138	5	1,703	University
Cora	19,793	126,842	70	8,710	Word/Degree
Arxiv	169,343	1,166,243	40	128	Time/Degree

Table 2: Statistics of Twitch-explicit for inductive tasks.

	DE	ES	FR	PTBR	RU	TW
Nodes	9,498	4,648	6,549	1,912	4,385	2,772
Edges	153,138	59,382	112,666	31,299	37,304	63,462
Density	0.003	0.006	0.005	0.017	0.004	0.017
Transit	0.047	0.084	0.054	0.131	0.049	0.057

Experiment

- Compared methods
  - **ERM:** Empirical Risk Minimization
  - **IRM:** Invariant Risk Minimization
  - **VREx:** Variance Risk Extrapolation
  - **GroupDRO:** minimizes the worst-case training loss over a set of pre-defined groups
  - **DANN:** adversarially trains the regular classifier and a domain classifier
  - **DeepCoral**: minimizes the deviation of covariant matrices from different domains
  - **Mixup:** a two-branch Mixup graph convolution to interpolate the irregular graph topology
  - **SRGNN**: Shift-Robust GNN
  - **EERM:** Explore-to-Extrapolate Risk Minimization
- Metrics
  - **Accuracy:** $\text{eval}(\mathbf{Y}, f_{\mathbf{W}}(\mathbf{A}, \mathbf{X}))$

Experiment

➢Evaluation and ablation under **covariate** shift

- FLOOD outperforms current state-of-the-art methods on OOD settings due to its flexibility during the test phase

Dataset		CBAS		WebKB		GOOD-Cora				GOOD-Arxiv			
Domain		Color		University		Word		Degree		Time		Degree	
Covariate		ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
Base	ERM	91.43	76.43	37.70	14.29	70.69	64.82	73.32	56.25	72.53	70.77	77.58	58.22
Invariant Learning	IRM	91.43	78.57	42.62	16.67	71.00	65.09	73.72	56.02	72.56	71.26	77.51	58.98
	VREx	92.86	78.57	36.07	16.67	70.79	64.77	73.32	56.28	72.58	71.25	77.73	58.95
	GroupDRO	92.86	80.00	42.62	14.29	70.74	64.82	73.32	56.37	72.61	71.14	77.63	59.09
Domain Generalization	DANN	94.29	74.29	42.62	15.25	70.69	64.77	73.32	56.20	72.48	71.16	77.38	59.19
	DeepCoral	91.43	78.57	39.34	16.67	70.69	64.80	73.32	56.34	72.76	71.28	77.75	59.20
Augmentation	Mixup	80.34	72.86	50.82	18.63	71.70	65.19	74.33	56.28	72.52	71.03	77.60	57.90
Graph OOD	SRGNN	87.14	71.43	42.62	13.32	70.14	64.32	71.00	53.88	72.25	70.77	76.05	57.66
	EERM	84.29	70.00	47.54	17.06	69.98	62.55	73.32	56.40	OOM	OOM	OOM	OOM
Ablation	FLOOD\Inv	90.26	75.78	42.31	15.34	70.23	63.23	73.23	56.32	72.41	70.21	77.42	56.82
	FLOOD\TtT	90.35	79.23	42.56	17.43	70.57	64.57	72.45	56.21	72.12	71.23	77.23	58.27
Ours	FLOOD	91.34	**83.53**	43.72	**18.95**	70.35	**66.23**	73.24	**56.64**	72.44	**72.13**	77.81	**59.47**

Experiment

➢Evaluation and ablation under **concept** shift

- FLOOD outperforms current state-of-the-art methods on OOD settings due to its flexibility during the test phase

Dataset		CBAS		WebKB		GOOD-Cora				GOOD-Arxiv			
Domain		Color		University		Word		Degree		Time		Degree	
Concept		ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
Base	ERM	90.00	81.43	63.33	26.61	65.90	64.35	69.00	61.09	74.86	67.35	75.06	62.29
Invariant Learning	IRM	90.71	82.52	61.67	27.23	65.96	64.40	68.04	61.23	74.37	67.40	75.38	62.49
	VREx	90.00	82.14	63.33	28.44	65.90	64.37	68.93	61.10	74.74	67.29	74.96	62.72
	GroupDRO	89.29	83.57	63.33	29.92	66.09	64.49	68.87	61.12	74.51	67.47	75.22	62.63
Domain Generalization	DANN	90.00	82.71	63.33	26.61	65.83	64.53	68.93	61.03	74.76	67.03	74.91	62.55
	DeepCoral	90.00	81.43	63.21	28.42	66.09	64.49	69.13	61.14	74.82	67.62	75.07	62.49
Augmentation	Mixup	93.57	64.29	63.33	30.28	70.58	64.77	70.15	63.12	74.74	65.17	72.28	60.10
Graph OOD	SRGNN	90.00	80.71	68.33	25.69	65.96	65.20	69.26	60.62	74.56	67.15	74.81	62.07
	EERM	81.43	62.14	63.33	26.53	65.06	62.66	65.85	58.23	OOM	OOM	OOM	OOM
Ablation	FLOOD\Inv	90.25	82.35	63.24	25.23	65.32	64.24	68.34	61.24	74.32	67.21	74.21	62.34
	FLOOD\TtT	90.32	82.31	63.56	28.35	65.82	64.46	68.23	61.23	74.21	67.12	74.23	62.52
Ours	FLOOD	90.47	**84.25**	63.72	**31.95**	65.85	**65.23**	68.24	**63.64**	74.24	**67.93**	74.81	**63.47**

Experiment

➢Evaluation and ablation under **inductive** distribution shift

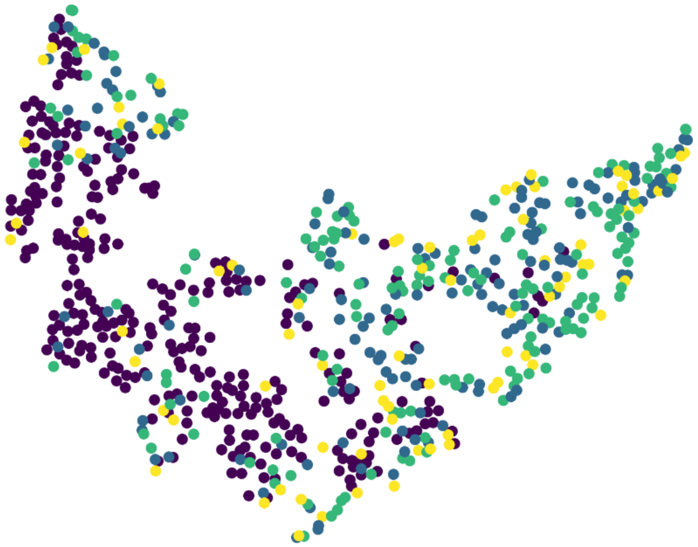
- The bootstrapped learning component in FLOOD leads to better generalization performance on inductive tasks than transductive tasks.

Dataset		ES		FR		PTBR		RU		TW	
Metric		AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc
ID		68.23	58.47	68.69	66.75	69.12	65.61	64.76	63.60	64.81	62.48
Base	ERM	62.10	45.59	62.12	42.41	62.60	50.71	50.25	38.52	51.29	40.95
Invariant Learning	IRM	63.50	49.97	62.74	43.53	63.69	51.85	55.19	39.33	51.90	41.47
	VREx	64.53	43.98	62.37	46.02	64.20	51.77	52.76	41.29	54.80	42.40
	GroupDRO	64.78	50.76	62.48	46.56	64.22	51.91	55.60	43.41	54.72	40.79
Domain Generalization	DANN	62.20	43.19	62.62	46.69	64.55	47.03	55.50	44.32	54.19	41.59
	DeepCoral	63.03	43.61	62.75	46.53	64.73	47.92	55.75	44.63	54.82	42.39
Augmentation	Mixup	62.28	47.07	60.95	40.92	61.73	46.81	54.76	35.63	56.98	44.24
Graph OOD	SRGNN	63.30	42.72	60.38	43.65	60.69	54.28	54.53	41.04	55.45	42.11
	EERM	65.18	51.74	63.04	46.86	64.91	51.49	56.68	44.91	58.77	46.07
Ablation	FLOOD\Inv	63.63	42.74	62.36	43.21	62.12	51.43	52.52	40.21	52.37	40.21
	FLOOD\TtT	64.32	43.84	63.45	46.23	64.23	52.32	53.25	42.53	55.21	42.93
Ours	FLOOD	**66.77**	**54.95**	**65.48**	**48.66**	**65.59**	**56.98**	**57.13**	**45.80**	**59.93**	**48.99**

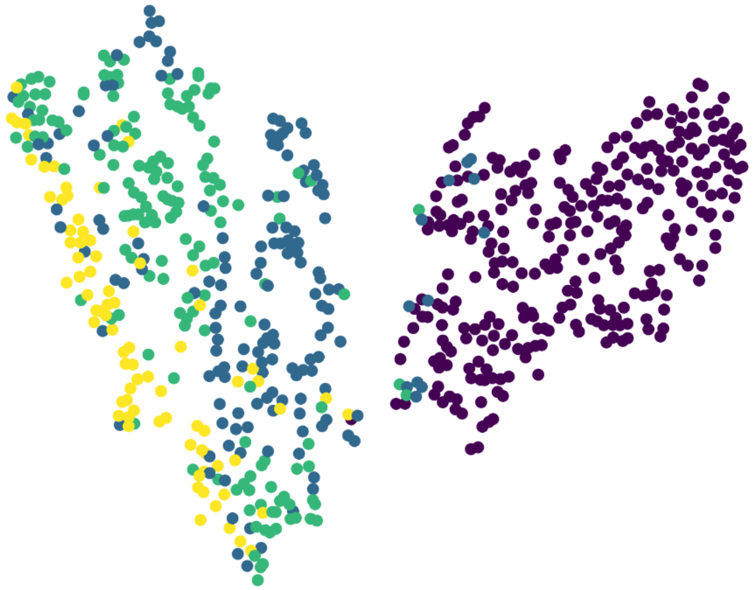
Experiment

➢RQ3: How does test-time training improve the generalization of GNNs?

- The shared encoder fine-tuned by FLOOD learns **more discriminative** representations, thanks to the bootstrapped learning during the test phase.



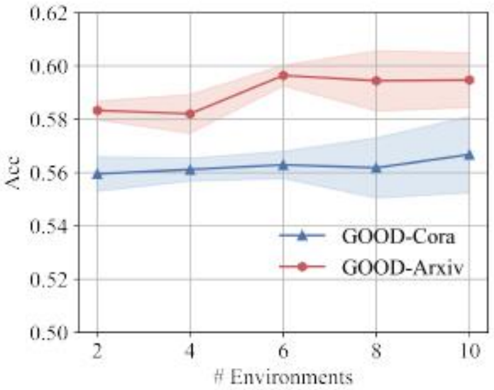
Before test-time training



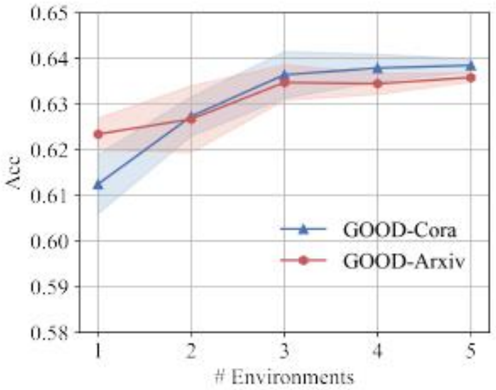
After test-time training

Experiment

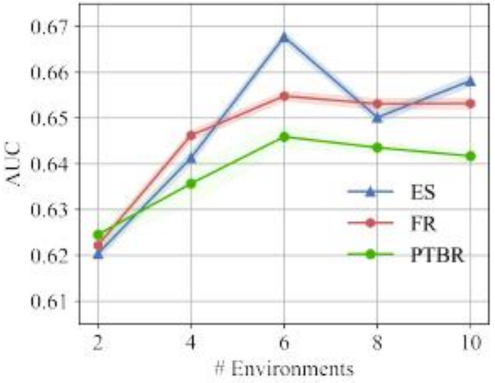
➢RQ4: What is the sensitivity of FLOOD with respect to the number of training environments and gradient descent steps during the test phase?



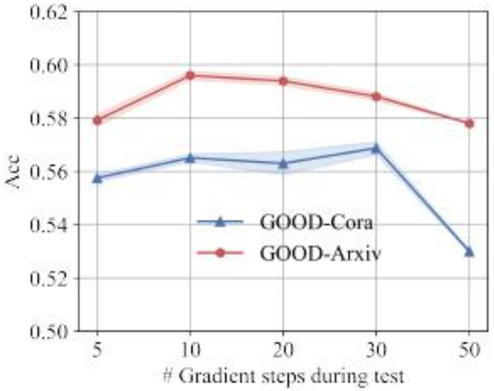
(a) Acc for covariate shift with different numbers of environments



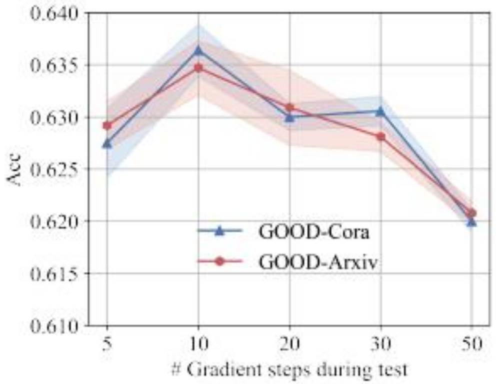
(b) Acc for concept shift with different numbers of environments



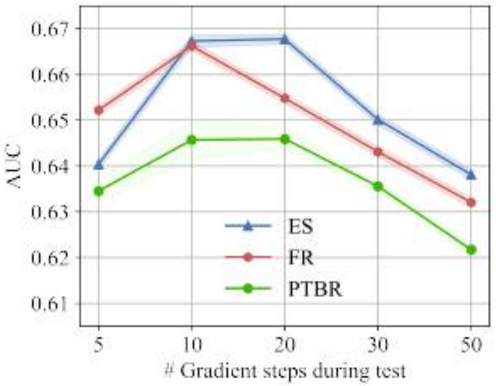
(c) Acc for inductive tasks with different numbers of environments



(d) Acc for covariate shift with different gradient steps during the test phase



(e) Acc for concept shift with different gradient steps during the test phase



(f) Acc for inductive tasks with different gradient steps during the test phase

Content

- Background and Motivation
- Method – FLOOD
- Experiment
- **Conclusion**

Conclusion

- We investigated the issue of out-of-distribution (OOD) generalization in graph representation learning.
- We proposed a new solution, FLOOD, which combines invariant representation learning and bootstrapped representation learning.
- FLOOD aims to find a balance between **stability** across different training environments and **adaptability** to the test distribution.
- Experiments on OOD benchmark graph datasets demonstrate the effectiveness of the proposed FLOOD framework.

Acknowledgment

Thanks for listening!

If you have any question, feel free to contact us at

liuyang520ict@gmail.com

heqing@ict.ac.cn

Paper and slides are available at

https://ponderly.github.io/