Towards a Theoretical Framework of Out-of-Distribution Generalization

Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, Liwei Wang

Peking University, Huawei Noah's Lab

NeurIPS 2021

Background and Motivation

- Out-of-Distribution (Domain) generalization
  Train on sets of available domains, and generalize to data sampled unseen domains.
- Why can we expect this generalization?
  Some assumption should be made in replace of i.i.d. Assumption.
- Immediate question
  - What assumption should we make?
  - Can this assumption capture the learnability?
  - How large the OOD generalization error gap can we get?

Our contribution:

- We formalize our assumption to OOD generalization and build a framework to describe its learnability.
- We prove upper and lower bound for generalization error gap.
- We design model selection method inspired by our bounds.

Problem Formalization

- Consider a classification task $\mathcal{X} \to \mathcal{Y} = \{1, ..., K\}$.
- An OOD generalization task is $(\mathcal{E}_{avail}, \mathcal{E}_{all})$, where $\mathcal{E}_{avail} \subset \mathcal{E}_{all}$ are domain sets.
- An OOD algorithm $\mathcal{A} : \mathcal{E}_{avail} \mapsto \mathcal{F}$, where $\mathcal{F}$ is the hypothesis space.
- Typically, $f \in \mathcal{F}$ can be decomposed into $g \circ h$: $g \in \mathcal{G}$ is the top model and

$$h(x) = (\phi_1(x), ..., \phi_d(x))^\top, \phi_i \in \Phi \text{ is feature.}$$

- OOD goal: minimize worse-domain loss, i.e., minimize

$$\mathcal{L}(f, \mathcal{E}_{all}) \triangleq \max_{e \in \mathcal{E}_{all}} \ell(f, e)$$

- OOD generalization error: $err(f) \triangleq \mathcal{L}(f, \mathcal{E}_{all}) - \mathcal{L}(f, \mathcal{E}_{avail})$

Our Assumption

A good assumption should:

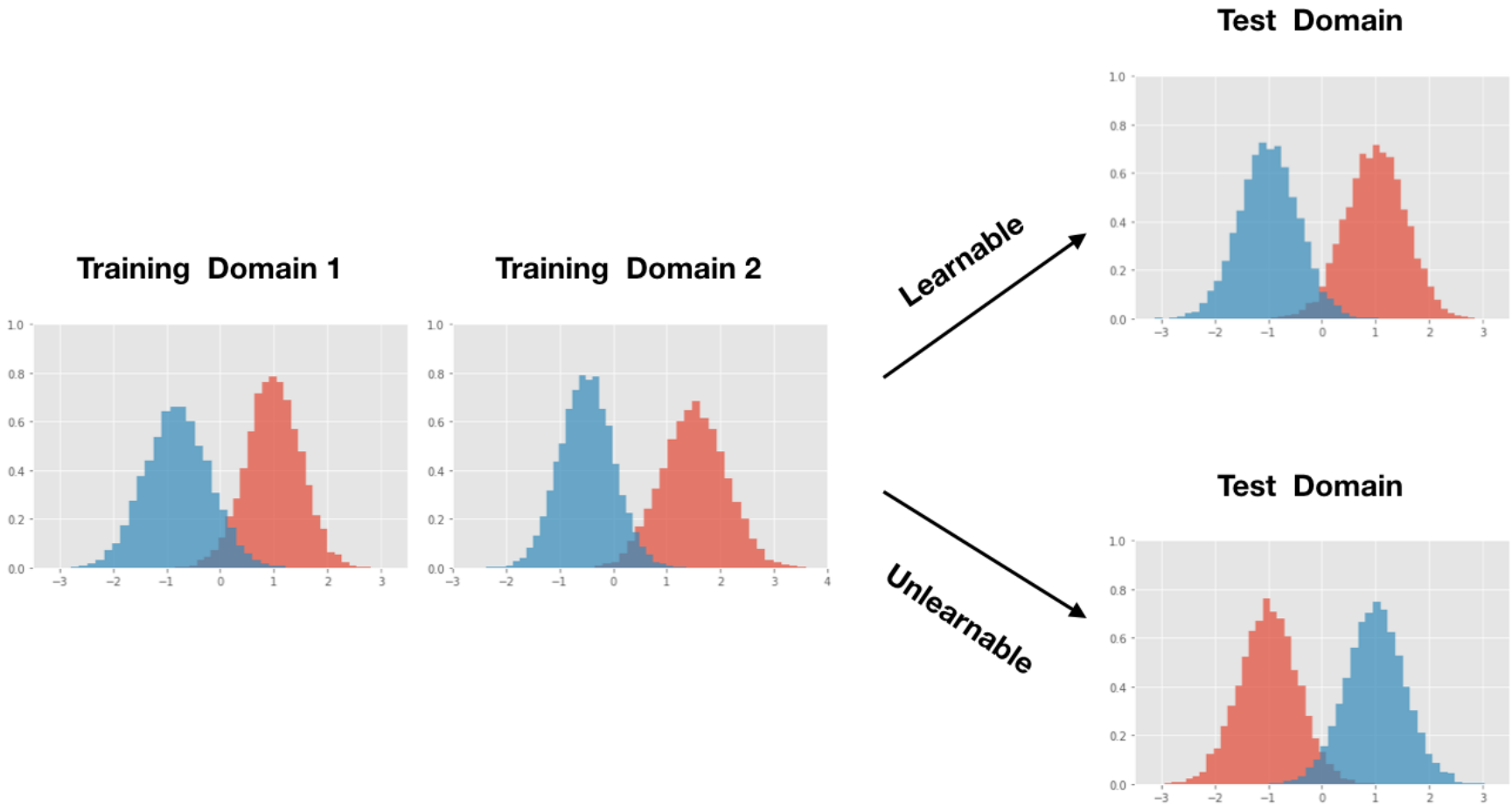
- Coincide with our intuition on what property an OOD generalization problem should have.
- Unify OOD generalization problem of different difficulty (learnability).

In our understanding:

- Assumption: the invariance of any informative feature ϕ across $\mathcal{E}_{avail}$ should be preserved in $\mathcal{E}_{all}$ to some extent.
- Difficulty: the larger this variation is amplified, the harder this OOD generalization task will be.
- Reason: if an informative feature is invariant across $\mathcal{E}_{avail}$, we have no way to refuse learning it. If it turns out to vary a lot in $\mathcal{E}_{all}$, it's unlikely to guarantee any generalization.

Illustration

Assumption: the invariance of any informative feature ϕ across $\mathcal{E}_{avail}$ should be preserved in $\mathcal{E}_{all}$ to some extent.



Variation and Informativeness

For a feature $\phi(\cdot)$, denote $P_y^e = \mathbb{P}(\phi(X^e|Y=y))$ as the distribution of $\phi(X)$ in domain e given $Y=y$.

Definition (Variation)

The variation of a feature $\phi(\cdot)$ across a domain set $\mathcal{E}$ is

$$\mathcal{V}(\phi,\mathcal{E}) = \max_{y\in\mathcal{Y}} \sup_{e,e'\in\mathcal{E}} \rho\big(P_e^y, P_{e'}^y\big)$$

Definition (Informativeness)

The informativeness of a feature $\phi(\cdot)$ across a domain set $\mathcal{E}$ is

$$\mathcal{I}(\phi,\mathcal{E}) = \frac{1}{K(K-1)} \sum_{y\neq y'} \min_{e\in\mathcal{E}} \rho\big(P_y^e, P_{y'}^e\big)$$

Learnability of OOD Generalization

Definition (Expansion Function)

We say a function $s : R^+ \cup \{0\} \to R^+ \cup \{0, +\infty\}$ is an expansion function, iff the following properties hold: 1) $s(\cdot)$ is monotonically increasing and $s(x) \geq x, \forall x \geq 0$; 2) $\lim_{x \to 0^+} s(x) = s(0) = 0$.

Let Φ be the feature space and ρ be a distribution distance.

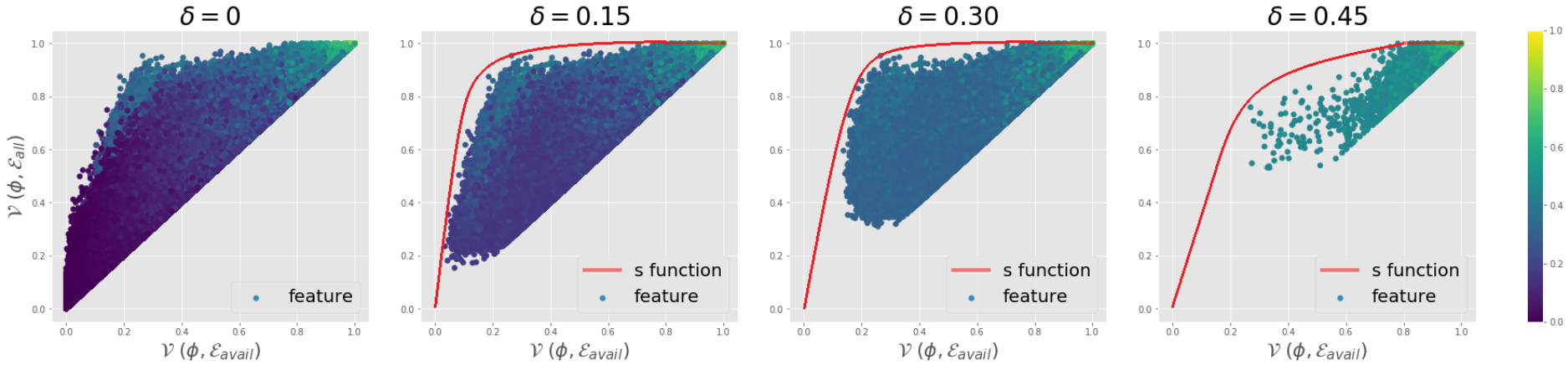
Definition (Learnability)

We say an OOD generalization problem from $\mathcal{E}_{avail}$ to $\mathcal{E}_{all}$ is $(s(\cdot), \delta)$-learnable if there exists an expansion function $s(\cdot)$ and $\delta \geq 0$, such that: for all $\phi \in \Phi$ satisfying $\mathcal{I}_\rho(\phi, \mathcal{E}_{avail}) \geq \delta$, we have $s(\mathcal{V}_\rho(\phi, \mathcal{E}_{avail})) \geq \mathcal{V}_\rho(\phi, \mathcal{E}_{all})$.

If an OOD generalization problem is not learnable, we call it unlearnable.

Expansion Function in Real Dataset

One may wonder what an expansion function is like in a real-world dataset Office-Home[1].



- We train thousands of models and obtain millions of features, with Φ chosen as ResNet-50.
- We set δ as different values, and see what the $s(\cdot)$ is like.

[1] https://arxiv.org/abs/1706.07522

OOD Upper Bound

Recall that $f = g(h(x)), h(x) = (\phi_1(x), ..., \phi_d(x))^\top$. Denote the random variable $h(X^e|Y = y)$ as $\mathbf{h}^e|y$, and the distribution of it as $p_{\mathbf{h}^e|y}(\cdot)$.

Theorem (General Case)

Under some assumption on the decay rate of $p_{\mathbf{h}^e|y}$ and its Fourier Transform $\hat{p}_{\mathbf{h}^e|y}$, if $(\mathcal{E}_{avail}, \mathcal{E}_{all})$ is $\left(s(\cdot), \mathcal{I}^{inf}(h, \mathcal{E}_{avail})\right)$-learnable, then

$$err(f) \leq O\left(s\left(\mathcal{V}^{sup}(h, \mathcal{E}_{avail})^{\frac{\alpha^2}{(\alpha+d)^2}}\right)\right).$$

Theorem (Linear Top Model)

Assume $\ell(\hat{y}, y) = \sum_{k=1}^{K} \ell_0(\hat{y}_k, y_k)$ and $g(x) = Ax + b$. If $(\mathcal{E}_{avail}, \mathcal{E}_{all})$ is $\left(s(\cdot), \mathcal{I}^{inf}(h, \mathcal{E}_{avail})\right)$-learnable, then

$$err(f) \leq O\left(s\left(\mathcal{V}^{sup}(h, \mathcal{E}_{avail})\right)\right).$$

Trade-off

To decrease the value of $s\left(\mathcal{V}^{\sup}(h, \mathcal{E}_{avail})\right)$, we can

- decrease $\mathcal{V}^{sup}(h, \mathcal{E}_{avail})$
- find a $s(\cdot)$ which increases slower.

But these two are contradictory in some way!

- We can realize the trade-off between learning informative feature and invariant feature.
- This trade-off coincides with our intuition.
- We also prove a lower bound of $\mathrm{err}(f)$, and design a model selection method based on our framework.

Takeaway

- We give our assumption on OOD generalization in replace of i.i.d. assumption.
- We mathematically formalize the assumption and construct our framework using informativeness, variation and expansion function.
- We prove a generalization bound, and give theoretical support on the importance of finding a balance between informativeness and variation.

This is Haotian Ye, and I'm going to apply PhD on next winter (2022), and please feel free to contact me if you find my work interesting!

my email: haotianye@pku.edu.cn

Thanks!