Out-of-distribution Generalization with Causal Invariant Transformations

Inspiration：能否想办法直接分出causal和non-causal feature，然后用causal feature训练？（或，在测试时，想办法分出causal feature并用其做预测）（本文给出的结果是，causal feature很难找，但是寻找使causal feature保持不变的变换较为容易）

Invariant Causal Mechanism

此类方法的主要假设是因果机制在不同domain不变。*Elements of causal inference: foundations and learning algorithms*.

其中一类方法：恢复causal structure。限制：linear structural model/足够多的domain。

但：[ICLR2022]Invariant causal representation learning for Out-of-Distribution Generalization提出了非线性情况下的

本文提出的不变因果机制：

$$Y = m(g(X), \eta),\ \eta \perp\!\!\!\perp g(X) \text{ and } \eta \sim F,$$

where X, Y are respectively the observed input and outcome, $g(X)$ denotes the causal feature, η is some random noise, and $m(\cdot,\cdot)$ represents the unknown structural function. The relationship $\eta \perp\!\!\!\perp g(X)$ means that the noise η is independent of the causal feature $g(X)$, and $\eta \sim F$ indicates that it follows a distribution F that can be unknown.

该机制中的spurious correlation：

1. X可能和噪声η相关联（尽管causal feature $g(x)$独立于η）
2. causal feature $g(X)$可能与背景相关联

spurious correlation会随着domain变化。

本文的创新点：

1. 现有的很多方法中采取的假设是：$g(X)$是线性的，且噪声是加性的。本文不需要这两个假设。
2. 不需要显式地学出$g(X)$，因此不需要处理可辨识性的问题。

Theorem 1： **Theorem 1.** *If $P_s \in \mathcal{P}$, then $\mathcal{H}_s \subseteq \mathcal{H}_*$.*

其中： $\mathcal{P} = \{P_{(X,Y)} \mid (X,Y) \sim P_{(X,Y)} \text{ under structural model (1)}\}$,

$h^*(\cdot) \in \mathcal{H}_* := \arg\min_h \sup_{P \in \mathcal{P}} \mathbb{E}_P[\mathcal{L}(h(X), Y)]$, 是最优模型；

$\mathcal{H}_s = \left\{ \phi \circ g \;\middle|\; \phi(w) \in \arg\min_z \mathbb{E}_{P_s}[\mathcal{L}(z, Y) \mid g(X) = w] \right\}$,是在$\mathcal{S}$上训练出的基于causal feature $g(X)$的最优模型。

Theorem 1的证明过程补充：

按原文记号，设噪声η的支撑集是$\mathcal{U}$

$$\begin{aligned}\mathbb{E}_Q[L(h(X),Y)|X=x] &= \int_{\mathcal{U}} L(h(x), m(g(x),\eta))p(x,\eta|x)d\eta \\ &= \int_{\mathcal{U}} L(h(x), m(g(x,\eta))p(x|x)p(\eta|x)d\eta \\ &= \int_{\mathcal{U}} L(h(x), m(g(x),\eta))p_\eta(\eta)d\eta\end{aligned}$$

第三个等号用到了x和η的独立性。

Theorem 1告诉我们：只要能找到causal feature，只用一个domain的数据就能学到最优模型。但是，显式地找出$g(X)$在实际中是很难的。不过，形状等特征不随旋转/翻转等变换变化。这种变化是比causal feature 好找的。

Theorem 2：

Theorem 2. *If $P_s \in \mathcal{P}$, then for $\mathcal{H}_s$ defined in Eq. (3)*

$$\mathcal{H}_s \subseteq \arg\min_h \sup_{T \in \mathcal{T}_g} \mathbb{E}_{P_s}[\mathcal{L}(h(T(X)), Y)].$$

说明了：只要知道不变变换集：$\mathcal{T}_g = \{T(\cdot) : (g \circ T)(\cdot) = g(\cdot)\}$，就能学出source domain上的最优预测器。

Theorem 2的等效形式：

Let $\mathcal{P}_{\text{aug}} = \{P_{(X',Y)} \mid (X,Y) \sim P_s,\ X' = T(X),\ T \in \mathcal{T}_g\}$, then we can rewrite the minimax problem in (4) as

$$\min_h \sup_{P \in \mathcal{P}_{\text{aug}}} \mathbb{E}_P[\mathcal{L}(h(X), Y)], \tag{5}$$

弄成(5)的意义：相比于 $h^*(\cdot) \in \mathcal{H}_* := \arg\min_h \sup_{P \in \mathcal{P}} \mathbb{E}_P[\mathcal{L}(h(X), Y)], \quad (2)$ ，(5)的sup条件更好实现，因为只需要在P_{aug}里边找max就行了。但是，计算P_{aug}的上确界计算量也很大。比如，旋转角度可以是0~360°的任何度数。

Causal Essential Set：是不变变换集的一个子集，满足对于$g(X)$相同的输入x_1、x_2，这个causal essential set里存在有限多的变换使得x_1经过这有限个变换后和x_2相同。

Definition 2 (Causal Essential Set). *For $\mathcal{I}_g \subseteq \mathcal{T}_g$, $\mathcal{I}_g$ is a causal essential set if for all x_1, x_2 satisfying $g(x_1) = g(x_2)$, there are finite transformations $T_1(\cdot), \cdots, T_K(\cdot) \in \mathcal{I}_g$ such that $(T_1 \circ \cdots \circ T_K)(x_1) = x_2$.*

Theorem 3：

Theorem 3. *If $P_s \in \mathcal{P}$, then for any $\mathcal{I}_g$ that is a causal essential set of $g(\cdot)$ and $\mathcal{H}_s$ defined in (3)*

$$\mathcal{H}_s = \arg\min_h \mathbb{E}_{P_s}[\mathcal{L}(h(X), Y)], \quad \text{subject to } h(\cdot) = (h \circ T)(\cdot),\ \forall T(\cdot) \in \mathcal{I}_g. \tag{6}$$