1. Out-of-distribution Generalization with Causal Invariant Transformations

1.1. problem

- 目前的机器学习方法的成功是基于i.i.d假设的
- 现有的利用未知的causal mechanism机制来处理OOD问题的方法**往往需要多个domain的数据或者多样的训练数据**,这使得该类方法在现实中应用受限

1.2. contributions

- 作者证明了若**已知全部的causal invariant transformations**(不改变causal feature只改变非causal feature的变换),可以**仅使用一个domain的数据学习一个具有OOD泛化能力的模型**.具体而言,**若全部的CIT已知,并施加在数据集之上,在经过transfermation之后的数据集上训练一个损失最小的模型,可以得到一个在所有的domain中拥有minimax optimality性能的模型**
- 作者证明了只需要知道一定合适的transfermation集合就可以完成OOD任务

1.3. assumption

- invariant causal mechanism assumption.

1.4. method

1.4.1. invaraint causal mechanism

数据生成过程使用下式来描述:

$$Y = m(g(X), \eta), \eta \perp\!\!\!\perp g(X) \text{ and } \eta \sim F \tag{1}$$

(这篇文章中使用的SCM更加的一般,所以方法的适用范围更广)其中$g(X)$代表**因果特征**(原数据经过映射后得到的特征).函数$m()$表示因果结构方程,η表示噪声,且服从分布F,$\eta \perp\!\!\!\perp g(X)$.这一结构中存在两个关联性:

1. 虽然$\eta \perp\!\!\!\perp g(X)$,但是$X \not\perp\!\!\!\perp \eta$可能是存在的
2. 因果特征$g(X)$与其他非因果特征之间可能存在相关性

这两类相关性可能会大大降低经验风险最小模型的效果. 例如分辨马与骆驼的任务中,causal feature是马或者骆驼的形状,但是马或者骆驼又往往与图片的背景是有关联的,所以训练出来的模型可能错误地仅仅以来图片的背景来对图片中的内容进行判断

1.4.2. 利用causal feature进行泛化(在已知causal feature的情况下的泛化模型的求解)

SCM(1)对应的数据分布如下

$$\mathcal{P} = \left\{ P_{(X,Y)} \mid (X,Y) \sim P_{(X,Y)} \text{ under structural model } (1) \right\} \tag{2}$$

因为SCM模型中可能只包含了一部分变量,所以$\mathcal{P}$中可能包含多种分布.我们的目标是

$$h^*(\cdot) \in \mathcal{H}_* := \arg\min_h \sup_{P \in \mathcal{P}} \mathbb{E}_P[\mathcal{L}(h(X), Y)], \tag{3}$$

即在各个$P \in \mathcal{P}$中最大损失最小的模型h^*.

令P_S代表源域数据,给定因果特征$g(X)$情况下的最优模型集合为

$$\mathcal{H}_s = \left\{ \phi \circ g \middle| \phi(w) \in \arg\min \mathbb{E}_{P_s}[\mathcal{L}(z, Y) \mid g(X) = w] \right\}, \tag{4}$$

其中$\circ$表示函数的嵌套.于是作者证明了

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

其中$\mathcal{H}_*$表示能够解决问题(3)的一系列模型集合.作者证明了当$g(X)$已知的情况下,可以仅使用源域数据P_S就可以得到$\mathcal{H}_*$的子集.

1.4.3. 利用causal invariant transformation进行学习

theorem1说明了如果已知causal feature $g(X)$,可以只使用一个domain的数据就可以学到最优的模型$\mathcal{H}_S$.但在**实际中可能并不知道$g(X)$的具体值**,并且使用$g(X)$的话其实还存在**可识别性**的问题.

作者在本文中提出来利用causal invariant transformation(定义如下,在原始特征上施加应用T后,causal feature不变)而不是causal feature的方案.

Definition 1 (Causal Invariant Transformation (CIT)). *A transformation $T(\cdot)$ is called a causal invariant transformation if $(g \circ T)(\cdot) = g(\cdot)$.*

令$\mathcal{T}_g = \{T(\cdot) : (g \circ T)(\cdot) = g(\cdot)\}$,为所有causal invariant transfermation的集合.当给定$\mathcal{T}$,$\mathcal{H}_S$也是可以求解的:

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

$$\mathcal{H}_s \subseteq \arg\min_h \sup_{T \in \mathcal{T}_g} \mathbb{E}_{P_s}[\mathcal{L}(h(T(X)), Y)]. \quad (4)$$

theorem说明当模型能在所有经过转化的数据上表现一致的好,那么就能在$\mathcal{P}$中的分布之间进行泛化.

令$\mathcal{P}_{\text{aug}} = \{P_{(X',Y)} \mid (X,Y) \sim P_s, X' = T(X), T \in \mathcal{T}_g\}$,可将theorem2修改为:

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

可以看到(5)与(3),很像,并且可以验证$\mathcal{P}_{aug}$是$\mathcal{P}$的子集.作者给出了以下两个remark

1. $\mathcal{P}_{aug}$是$\mathcal{P}$的真子集.(5)相较于(3)更好求解.(**作者举的例子属实无法理解**)
2. 使用马与骆驼的案例来说明,如果我们更换图片的背景,比如骆驼的背景即有可能是沙漠也有可能是草地,那么那种依赖于背景这种虚假特征的模型就没办法表现的很好,自然也不会是(5)的解

但仅适用causal invaraint transformation来进行学习计算量可能会很大,比如图像旋转的角度可能就有360中情况,计算这360种情况的loss可能计算量非常大

1.4.4. 利用Causal Essential Set进行学习

这一部分作者证明了**仅需要$\mathcal{T}$中的一部分或者说子集就可足够了**.
$\mathcal{T}_g$中包含了全部能让原始特征变换但casual feature不变的转化,而causal essential set中包含的是能让所有causal feature一致的数据可以相互转化的transformation:

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 $(T_1 \circ \cdots \circ T_K)(x_1) = x_2$.*

在给定任何causal essential set的一些先验知识后,能够达到OOD泛化能力:

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

$$\begin{aligned} \mathcal{H}_{\mathrm{s}} = \arg\min_{h} \ & \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(X), Y)], \\ \text{subject to } \ & h(\cdot) = (h \circ T)(\cdot), \ \forall\, T(\cdot) \in \mathcal{I}_g. \end{aligned} \tag{6}$$

相较于theorem2中的$\subset$,这里改成了$=$,是一个更强的理论结果.

1.4.5. 先验知识的必要性

需要从观测数据中获得causal results,那么先验知识是必要的.

目前一些现有的causal discovery的方法往往假设causal graph满足某种特定的形式因此在实际数据中很难应用.而另外的一些方案例如IRM,往往需要足够多的多个域的训练数据或者要对causal图的形式有一定的先验知识.

所以总的来说,只有通过随机实验的方法才是推断因果关系的黄金准则——“no causation without manipulation”

另外直接使用人工修改的数据也可以在提升因果估计的效应。但是人工的操作可能过于繁琐，作者认为如果使用一些先验知识来获得CIT比直接在数据上进行人为的操作更加的简单。

1.5. algorithm

引入距离算子$D()$,$D(v_1, v_2)$表示v_1, v_2之间的距离.若$v_1 = v_2$则$D(v_1, v_2) = 0$,于是theorem6中的式子可以替换为

$$\begin{aligned} & \min_{h} \mathbb{E}_{P_s}\left[\mathcal{L}(h(X), Y)\right], \\ \text{subject to} & \mathbb{E}_{P_s}\left[\sup_{T \in \mathcal{I}_g} D\big(h(X), h(T(X))\big)\right] = 0, \end{aligned}$$

转化为最小化:

$$\mathbb{E}_{P_s}[\mathcal{L}(h(X), Y)] + \lambda_0 \mathbb{E}_{P_s}\left[\sup_{T \in I_g}[D(h(X), h(T(X)))\right] \tag{6}$$

对于有限的训练集合,等价于最小化下式:

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(h(x_i),y_i)+\frac{\lambda_0}{n}\sum_{i=1}^{n}\sup_{T\in I_g}[D(h(x_i),h(T(x_i)))]. \tag{7}$$

算法的截图如下

Algorithm 1 Regularized training with Invariance on Causal Essential set (RICE).

Input: Training set $\{(x_1,y_1),\cdots,(x_n,y_n)\}$, batch size S, learning rate η, training iterations N, model $h_\beta(\cdot)$ with parameter β, initialized parameter β_0, regularization constant λ_0, causal essential set $\mathcal{I}_g$, and discrepancy measure $D(\cdot,\cdot)$.

1: **for** $i=1,\ldots,n$ **do**
2: Generate transformed samples $\{T(x_i)\}_{T\in\mathcal{I}_g}$.
3: **end for**
4: **for** $t=0,\cdots,N$ **do**
5: Randomly sample a mini-batch $\mathcal{S}=\{(x_{t_1},y_{t_1}),\cdots,(x_{t_S},y_{t_S})\}$ from training set.
6: Fetch the transformed samples $\{T(x_{t_1})\}_{T\in\mathcal{I}_g},\cdots,\{T(x_{t_S})\}_{T\in\mathcal{I}_g}$.
7: Update model parameters via first-order method e.g., stochastic gradient descent:

$$\beta_{t+1}=\beta_t-\frac{\eta}{S}\sum_{i=1}^{S}\nabla_\beta\mathcal{L}(h_\beta(x_{t_i}),y_{t_i})\Big|_{\beta=\beta_t}+$$
$$\eta\nabla_\beta\Big\{\frac{\lambda_0}{S}\sum_{i=1}^{S}\sup_{T\in\mathcal{I}_g}D\big(h_\beta(x_{t_i}),h_\beta(T(x_{t_i}))\big)\Big\}\Big|_{\beta=\beta_t}.$$

8: **end for**

1.6. cycleGAN

有些情况下,我们不清楚具体的CIT,但是我们有CIT前后的数据,那么就可以使用cycle gan来表征对应的生成关系.只要把源域的数据输入,cyclegan能给出经过CIT后的数据.cyclegan的内容可以参考

1.7. 一些想法

1. 从SCM中获得的分布$\mathcal{P}$到底是怎么样的
2. 这篇论文最核心的问题是如何获得CIT呢?如果不是简单的图像识别,在其他领域这种类似的先验知识很难获得.就比如在电网中,得到一个数据集之后,我们其实很难直接的了解到改变那些特征是不会让其对应的标签改变的.
3. 这篇文章的主要内容与数据增强其实是一致的,那么文章的主要贡献在与从causal invaraint的角度说明了只使用有限的CIT就可以得到能在各个domain上表现出色的模型,即换了个角度来证明了为啥数据增强有意义