Causality Inspired Representation Learning for Domain Generalization

problem

Domain generalization问题(本质是一个OOD问题),DG问题的**主流方法是通过统计模型来建立数据域标签之间的依赖关系,并尝试学到域Domain无关的表征**

model assumption

1. 每个输入特征都是由一系列的因果因素与非因果因素混合而成的

related work

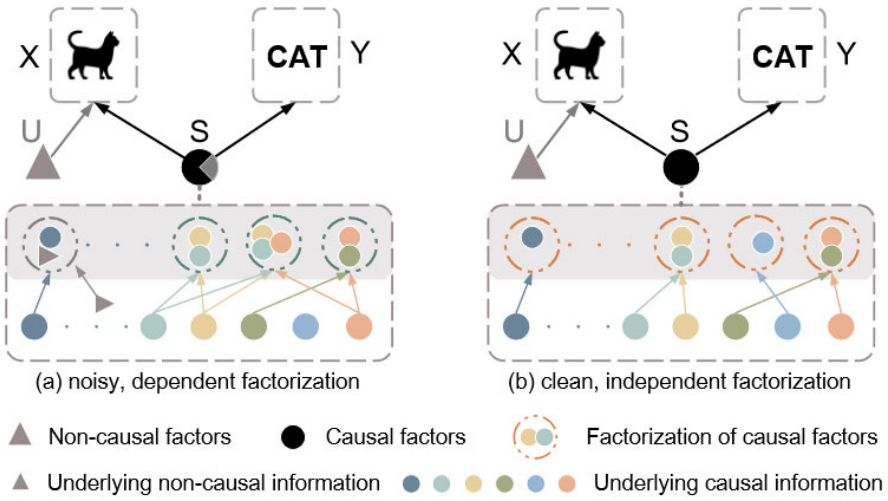
1. 为提高模型的泛化能力,有许多类别的方法,例如 `invariant representation learning` , `domain augmentation` , `meta-learning` 这类的方法还是局限于学习label与feature之间的相关性,而忽视了两者之间的causal机制
2.

Contribution

作者认为causal factors需要满足以下三个要求:

1. 能与非因果因子U分离(考虑后续的分类任务)
2. S的因式分解之间应该独立(**冗余方面的考虑**)
3. 包含所有的$X \rightarrow Y$的因果信息

以下图作为对比:



其中(b)中的S,就属于满足上述三条准则的causal factors 因为他对应的因子分解完全独立不存在冗余,并且能与U完全分离

基于此作者提出了Causality Inspired Representation Learning (CIRL).该模型的展开也围绕上述的三个要求,首先通过来干扰Domain相关的信息来生成新的数据,从而将S与U分离;接着,设计分解模块,使得S中的表征的每个维度都独立;最后使用对抗mask模块来提升S中某些维度的因果信息的含量.

Method

直接从原始数据中恢复causal factor是很困难的,因为causal factor往往是无法观测的.那么作者尝试基于causal factor的某些性质学习causal representation来作为causal factor的代替.

从causal的视角来看DG问题

causality与statistic dependence之间的关联,可表示为:

Principle 1 ([54]). ***Common Cause Principle***: *if two observables X and Y are statistically dependent, then there exists a variable S that causally influences both and explains all the dependence in the sense of making them independent when conditioned on S.*

结合principle1,我们可以构建以下的SCM来描述DG问题:

$$X := f(S, U, V_1), S \perp U \perp V_1, \tag{1}$$
$$Y := h(S, V_2) = h(g(X), V_2), V_1 \perp V_2. \tag{2}$$

其中,S, U分别代表causal factor与非causal factor,V表示噪声.示意图如下:

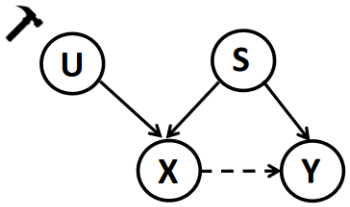


Figure 1. SCM of DG. The **solid arrow** indicates that the parent node causes the child one; while the **dash arrow** means there exists statistical dependence.

那么由上式可以得到,对于任意的分布$P(X, Y) \in \mathcal{P}$,给定S时.$P(Y|S)$是不变的(因为不管分布如何变化,(2)式一直是稳定存在的).于是我们就可以获得函数h:

$$h^* = \arg\min_h \mathbb{E}_P[\ell(h(g(X)), Y)] = \arg\min_h \mathbb{E}_P\big[\ell(h(S), Y)\big], \tag{3}$$

但实际中我们很难获取S,虽然有一些工作指出,可以利用一些特定的分布转换以及监督信号以及特定的intervention来提取causal factor.但作者更加关注causal factor要满足的某些requirements,例如ICM准则:

Principle 2 ([51, 58]). ***Independent Causal Mechanisms (ICM) Principle***: *The conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.*

那么当$S = \{s_1, s_2, ..., s_N\}$,改变$P(s_i|PA_i)$不会影响$P(s_j|PA_j)$,于是可以有以下的分解:

$$P(s_1, s_2, \ldots, s_N) = \prod_{i=1}^{N} P(s_i \mid PA_i), \tag{4}$$

针对principle1与principle2的性质,作者对causal representation,做出了以下的三个要求(使causal representation更加接近causal factor):

- The causal factors S should be separated from the non-causal factors U, i.e., $S \perp\!\!\!\perp U$. Thus, performing an intervention upon U does not make changes to S.
- The factorization $s_1, s_2, \ldots, s_N$ should be jointly independent, none of which entails information of others.
- The causal factors S should be causally sufficient to the classification task $X \to Y$, i.e., contain information that can explain all the statistical dependencies.

基于因果性的表征学习

causal intervention module

这部分作者通过对U施加intervention 来生成多个不同的Domain的数据(这里通过对幅值施加干预来实现对),并且利用先验知识$P(S|do(U))$保持不变来寻找causal factor S.

于是作者利用了傅里叶变化,傅里叶谱的相位分量保留了原始信号的高级语义，而幅度分量包含低级统计数据

给定原有的输入x^o,使用傅里叶变化得到

$$\mathcal{F}(x^o) = \mathcal{A}(x^o) \times e^{-j \times \mathcal{P}(x^o)}, \tag{5}$$

作者使用原数据x^o以及来自任意训练分布的数据$(x')^o$来打乱幅值信息:

$$\hat{\mathcal{A}}(x^o) = (1-\lambda)\mathcal{A}(x^o) + \lambda \mathcal{A}((x')^o), \tag{6}$$

接着将得到的经过扰动后的幅值信息与原相位信息结合,使用傅里叶逆变换得到数据x^a:

$$\mathcal{F}(x^a) = \hat{\mathcal{A}}(\boldsymbol{x}^o) \times e^{-j \times \mathcal{P}(x^o)}, \boldsymbol{x}^a = \mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{x}^a)). \tag{7}$$

于是训练一个特征提取器$\hat{g}() \in \mathbb{R}^N$,要求在输入为$x^o$与$x^a$是结果比较一致(用相关系数来表示):

$$\max_{\hat{g}} \frac{1}{N} \sum_{i=1}^{N} COR(\tilde{r}_i^o, \tilde{r}_i^a), \tag{8}$$

其中$\tilde{r}_i^o, \tilde{r}_i^a$分别是经过Z标准化后的$\check{\mathbf{R}}^o = [(\check{r}_1^o)^T, \ldots, (r_B^o)^T]^T \in \mathbb{R}^{B \times N}, \mathbf{R}^a = [(\boldsymbol{r}_1^a)^T, \ldots, (\bar{r}_B^a)^T]^T$的第i列.其中$B$为batchsize.

使用相关系数COR表示两者的相关性,两者相关性越大,就说明可以从U中分离出不会改变的causal representation R

Causal Factorization Module

作者对causal representation的第二个要求是各个维度之间相互独立.于是可以用各个causal representation之间的相互独立性来表征:

$$\min_{\hat{g}} \frac{1}{N(N-1)} \sum_{i \neq j} COR(\tilde{r}_i^o, \tilde{r}_j^a), i \neq j, \tag{9}$$

为了简便,作者省略了$\mathbf{R}^o, \mathbf{R}^a$内部的约束(**我觉得是为了方便后续推导**)
使用(7)(8),可以直接写成一个协方差矩阵的形式:

$$\boldsymbol{C}_{ij} = \frac{< \tilde{r}_i^o, \tilde{r}_j^a >}{||\tilde{r}_i^o|| ||\tilde{r}_j^a||}, i, j \in 1, 2, \ldots, N, \tag{10}$$

于是目标函数可以设计为

$$\mathcal{L}_{Fac} = \frac{1}{2} \|\boldsymbol{C} - \boldsymbol{I}\|_F^2. \tag{11}$$

同时约束各个causal representation能够在干预下保持稳定,同时彼此之间又能没有冗余重合部分.

Adversarial Mask Module

为了提升各个维度中包含的因果信息,作者尝试将识别包含信息较少的维度,并增加全新信息(要保持各个维度之间独立),使其对分类任务有更大的贡献值

最直接训练的方法是利用监督标签y,在多个Domain中间进行训练:

$$\mathcal{L}_{cls} = \ell(\hat{h}(\hat{g}(\boldsymbol{x}^o)), y) + \ell(\hat{h}(\hat{g}(\boldsymbol{x}^a)), y) \tag{12}$$

但这种训练模式没办法保证每个causal representation的维度都能包含充足的信息,可能存在一部分欠优的representation对于分类任务的贡献很小.

为解决该问题,作者引入了基于NN的mask层$\hat{\omega}$来学习每个causal representation的贡献值,并选择$\kappa N \in$个作为出色的causal representation或者说包含信息较多的representation,其中$\kappa \in (0, 1)$:

$$m = \text{Gumbel-Softmax}(\hat{w}(r), \kappa N) \in \mathbb{R}^N, \tag{13}$$

使用Gum本来 softmax trick是为了采样的操作可微分.

针对superior与inferior分别设计两个分类器:

$$\begin{aligned} \mathcal{L}_{cls}^{sup} &= \ell(\hat{h}_1(\boldsymbol{r}^o \odot \boldsymbol{m}^a), y) + \ell(\hat{h}_1(\hat{r}^a \odot \boldsymbol{m}^a), y), \\ \mathcal{L}^{inf} &= \ell(\widehat{h}_2(\boldsymbol{r}^o \odot (1 - \boldsymbol{m}^o)), y) + \ell(\hat{\boldsymbol{h}}_2(\hat{r}_2(\boldsymbol{r}^a \odot (\boldsymbol{1} - \boldsymbol{m}^a)), y), \end{aligned} \tag{14}$$

其中mask的目的是拆选出真正包含更多信息的causal representation,于是mask的目标是为令$\mathcal{L}_{cls}^{SUP}$最小,同时令$\mathcal{L}^{inf}$最大.而分类器h_1, h_2以及生成器g的目标是让模型预测更加准确,即最小化$\mathcal{L}_{cls}^{SUP}$与$\mathcal{L}^{inf}$.于是模型的的目标函数可以总结为一个adversarial的形式:

$$\min_{\hat{g}, \hat{h}_1, \hat{h}_2} \mathcal{L}_{cls}^{sup} + \mathcal{L}_{cls}^{inf} + \tau \mathcal{L}_{Fac} \quad \min_{\hat{w}} \mathcal{L}_{cls}^{sup} - \mathcal{L}_{cls}^{inf}, \tag{15}$$

Experiment

可解释性的测试

作者验证了使用CIRL学习得到的causal factor的可解释性以及合理性

想法

1. 这篇文章是假设有一个更深层的causal factor在影响X与Y,从而找到S,既可以获得稳定的$P(Y|S)$,我之前的工作其实都还是局限于在原始特征X上找能令$P(Y|X)$不变的特征,相当于他的假设比我更加的宽泛.
2. 其实(4)是由markov assumption得到的,并不是由ICM假设的得到的.
3. 作者提出的三个准则中,准则1与3是为了迎合principle1,准则2是为了迎合principle2
4. causal intervention module的设计想法其实和nonlinearICP的想法有一点不同,nonlinearICP是找到能让Y与E之间独立的特征,而causal intervention module是为了找到在各种intervention下$P(S|do(U))$不变的特征.不过我觉得这两者是可以转化的.
5. 傅里叶变换为何会具有这种特点呢?
6. (4)只是说明causal factor之间的机制是独立的,那么不代表各个维度本身是独立的
7. 作者测试模型的Domain generation的能力并不是只选用了一种未见过的Domain而是选择了多种Domain,再取平均的操作,这种方法很值得借鉴
8. 消融实验用于测试模型中各个模块缺失情况下,对模型表现的影响