一切诸果，皆从因起，一切诸报，皆从业起。在未学习因果理论之前，大家就知道世间万物皆有因果，我们是处在一个充满因果的复杂世界中。正是因为因果的存在，我们才能通过变化万千的事物总结规律形成自己的知识，做到见微知著，一叶知秋。这些都是人类通过学习可以掌握的能力，但是人工智能却很难通过训练模型来实现。虽然人工智能在经历无数次浪潮之后，在深度学习时代迎来巨大爆发，涌现出许多惊艳之作，打败人类围棋顶级选手的alphaGo、可以以假乱真的“下一个伦勃朗”项目、“只因在人群中多看你一眼”的天眼系统、基于大数据技术的新冠疫情预测、溯源、管控等。但是，我们必须承认深度学习并非万能，现在深度学习可以为机器人赋予强大的感知能力，但是在认知、决策层面仍力不从心。在深度学习模型的训练中，好的结果与海量优质标注数据集、精巧的网络结构、精心调整的超参数是密不可分的，而且针对某一领域或某问题学习到的模型无法直接泛化到相似的任务中去。很多研究者意识到造成这个现象的发生，一定程度上是由于模型缺乏逻辑推理能力，所以现在他们也开始在深度学习中融入因果的思想和理论。

深度学习方法建模因果推理CEVAE

CEVAE模型是首个使用VAE来建模因果推理的深度学习模型，作者通过引入医学治疗与患者健康之间因果关系的例子，假设了一幅包含treatment、outcome、confounder的因果图。但是，由于confounder在大多数情况下无法被准确的确定和观测，所以作者在原因果图的基础上添加了一个关于confounder的noisy views作为proxy，如图所示。

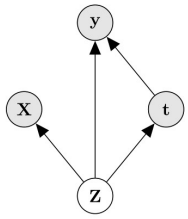


Figure 1: Example of a proxy variable. $\mathbf{t}$ is a treatment, e.g. medication; $\mathbf{y}$ is an outcome, e.g. mortality. $\mathbf{Z}$ is an unobserved confounder, e.g. socio-economic status; and $\mathbf{X}$ is noisy views on the hidden confounder $\mathbf{Z}$, say income in the last year and place of residence.

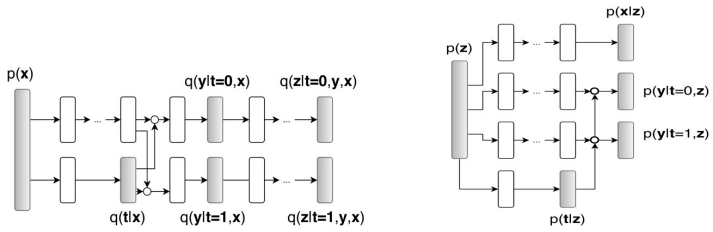
目前因果推理的研究主要就是集中在CATE或ATE的估计上，所以作者的目的就是建模ITE，从而得到ATE的估计。

$$ITE(x) := \mathbb{E}[y|X=x, do(t=1)] - \mathbb{E}[y|X=x, do(t=0)]$$
$$ATE(x) := \mathbb{E}[ITE(x)]$$

> ITE (Individual Treatment Effect): 评价treatment对于单个个体的效果。
>
> ATE (Average Treatment Effect): 评价treatment对于单个个体的效果。

$$\begin{aligned} p(y|X, do(t=1)) &= \int_Z p(y|X, do(t=1), Z) p(Z|X, do(t=1)) dZ \\ &\overset{2}{=} \int_Z p(y|X, t=1, Z) p(Z|X, do(t=1)) dZ \\ &\overset{3}{=} \int_Z p(y|X, t=1, Z) p(Z|X) dZ \\ &\overset{1}{=} \int_Z p(y|t=1, Z) p(Z|X) dZ \end{aligned} \tag{1}$$

从图中我们可以看出X是无法作为普通的混杂直接处理的，因为它与t和y之间没有直接的因果路径，直接处理它将会带来偏差。为了消除这种偏差，通过可观测的代理恢复真实的因果关系是解决问题的关键。**所以作者使用VAE学习一个潜在变量模型，用来发现隐藏的confounder并推理它们对t和y的影响。**虽然目前没有一个理论可以证明VAE可以学习真实的模型，但是它对数据生成和隐藏confounder的假设相较于其它方法很弱。



(a) Inference network, $q(\mathbf{z}, t, y|\mathbf{x})$. (b) Model network, $p(\mathbf{x}, \mathbf{z}, t, y)$.

Figure 2: Overall architecture of the model and inference networks for the Causal Effect Variational Autoencoder (CEVAE). White nodes correspond to parametrized deterministic neural network transitions, gray nodes correspond to drawing samples from the respective distribution and white circles correspond to switching paths according to the treatment t.

图a是对$q(z,t,y|x)$的建模，该模型参考的是另一个经典的casual inference深度模型TARnet，图b是对$p(x,z,t,y)$的建模，这个结构可以根据图1的因果图分解得到：

$$p(x,t,y,z) = p(z)p(t,x,y|z) = p(z)p(t,y|z)p(x|z) = p(z)p(t|z)p(y|t,z)p(x|z) \tag{2}$$

最终模型学习的目标函数为：

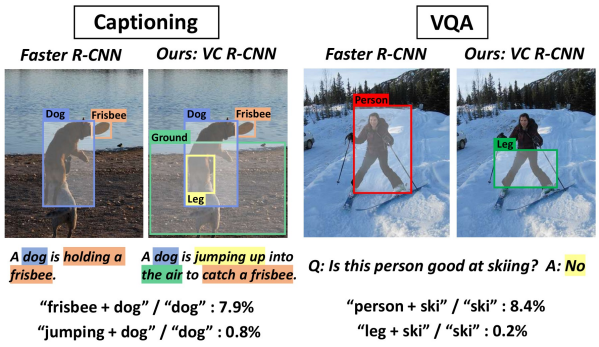
$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^{N} \mathbb{E}_{q(z_i|x_i,t_i,y_i)}[logp(x_i,t_i|z_i) + logp(y_i|t_i,z_i) + logp(z_i) - logq(z_i|x_i,t_i,y_i)] \\ &= \sum_{i=1}^{N} \mathbb{E}_{q(z_i|x_i,t_i,y_i)}[logp(x_i,t_i,y_i|z_i)] - \mathbb{E}_{q(z_i|x_i,t_i,y_i)}[log(q(z_i|x_i,t_i,y_i)/p(z_i))]\end{aligned} \tag{3}$$

$$\mathcal{F}_{\text{CEVAE}} = \mathcal{L} + \sum_{i=1}^{N}(logq(t_i = t_i^*|x_i^*) + logq(y_i = y_i^*|x_i^*, t_i^*)) \tag{4}$$

具体实现细节：整个网络结构主要由多个ELU非线性隐藏层网络组成，分别学习潜在变量Z的分布$q(Z|X,t,y)$，生成模型$p(X|Z)$，outcome模型$p(y|t,Z)$和treatment模型$p(t|Z)$、$q(t|X)$。为了计算最后的ITE结果，作者采用100多个样本的均值近似后验分布$q(Z|X) = \sum_t \int q(Z|t,y,X)q(y|t,X)q(t|X)dy$。

基于干预的图像常识特征VC R-CNN

目前深度学习模型具有较强的感知能力，在图像分割、图像分类、目标检测、目标跟踪等领域有着非常显著的成就，　但是缺乏认知能力，在更高级的应用层面如图像标注、视觉问答等领域效果仍有很大提升空间。**作者认为深度学习可以精确地回答what和where，却无法回答why，究其原因是因为机器缺乏常识而犯下“认知错误”。**对此，作者通过对比实验验证了这一结论。从图中我们可以看到在图像标注任务里缺乏常识的特征往往无法准确的描述视觉关系；在视觉问答任务里尽管模型可以做出正确的回答，但却将注意力放在了错误的视觉特征上。



此外，作者指出在NLP领域可以通过一个单词预测其上下文内容，在视觉领域通过学习局部特征却很难满足下游任务的需求，主要原因在于文本特征本身包含了常识，而图像在拍摄之后就无法观察物体的语境信息，所以对象间存在的常识就容易被虚假的观察偏移所混杂。为了说明这一点，作者在COCO数据集上对所有对象分别计算$P(Y|X)$和$P(Y|do(X))$来统计两者的差异，发现由于数据集的分布bias导致sink和hari drier条件概率远高于do概率，person和toilet的条件概率远小于do概率。

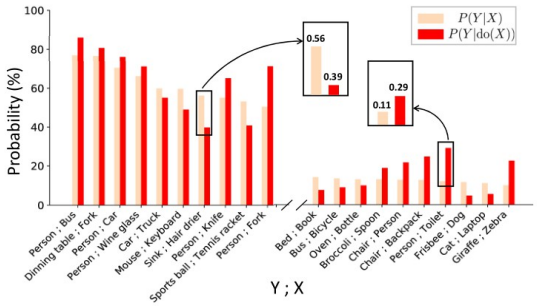


Figure 3. The sensible difference between the likelihood before (*i.e.*,$P(Y|X)$) and after intervention (*i.e.*, $P(Y|do(X))$) in MS-COCO. The object is represented by the 80 ground-truth class labels. Only 20 pairs are visualized to avoid clutter.

基于以上的实验启发，作者将intervention应用在目标检测任务中，并为intervention设计了一个proxy task：给定RoI X的特征去预测RoI Y的类别。作者在COCO数据集上将所有图像的object RoI特征在每个类别维度取平均，作为该类别的表示，进而构建出一个$N \times 1024$的confounder字典Z，其中包含着所有可能的混杂因子，然后采用如下图所示的intervention策略，利用backdoor理论消除confounder对X和Y的影响，由此得到一个鲁棒性更强的预测模型。

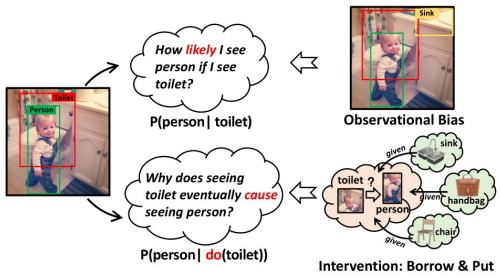


Figure 2. The illustration of why $P(Y|do(X))$ learns common sense while $P(Y|X)$ does not. Thanks to intervention, $P(Y|do(X))$ can "borrow" objects from other images and "put" them into the local image, to perform further justifications if X truly causes Y regardless of the unobserved confounders, and thus alleviate the observational bias.

作者将整个intervention整合成context predictor，同时添加self predictor保留了网络识别RoI X本身类别的能力，最终在Faster R-CNN的基础上构造了VC R-CNN，框架如图所示。

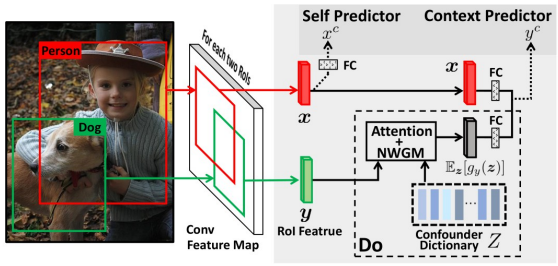


Figure 4. The overview of VC R-CNN. Any R-CNN backbone (*e.g.*, Faster R-CNN [54]) can be used to extract regions of interest (RoI) on the feature map. Each RoI is then fed into two sibling branches: a **Self Predictor** to predict its own class, *e.g.*, x^c, and a **Context Predictor** to predict its context labels, *e.g.*, y^c, with our **Do** calculus. The architecture is trained with a multi-task loss.

目标函数：

$$
\begin{aligned}
P(Y|do(X)) &= \sum_z P(y^c|x,z)P(z) \\
&= \sum_z Softmax(f_y(x,z))P(z) \\
&:= \mathbb{E}_z[Softmax(f_y(x,z))] \\
&\overset{NWGM}{\approx} Softmax(\mathbb{E}_z[f_y(x,z)])
\end{aligned} \tag{5}
$$

结合因果图分析，作者认为f_y是一个与x和z都相关的线性映射函数，其中g_z是名为Saled Dot-Product Attention的注意力机制，用来根据特定的y为混杂字典中的confounder赋予不同的权重值。

$$\mathbb{E}_z[f_y(x,z)] = W_1 x + W_2 \cdot \mathbb{E}_z[g_y(z)] \tag{6}$$

$$\mathbb{E}_z[g_y(z)] = \sum_z [Softmax((W_3 y)^T (W_4 Z^T)/\sqrt{\sigma} \odot Z)] P(z) \tag{7}$$

Index	Input	Operation	Output	Trainable Parameters
(1)	-	RoI feature	$\boldsymbol{x}$ (1024×1)	-
(2)	-	RoI feature	$\boldsymbol{y}$ (1024×1)	-
(3)	(2), $\boldsymbol{Z}$	Scale Dot-Product Attention	$\mathbb{E}_{\boldsymbol{z}}[g_y(\boldsymbol{z})]$ (1024×1)	$\boldsymbol{W}_3$ (512×1024) $\boldsymbol{W}_4$ (512×1024)
(4)	(1),(3)	Linear Addition Model	$\mathbb{E}_{\boldsymbol{z}}[f_y(\boldsymbol{x},\boldsymbol{z})]$ (80×1)	$\boldsymbol{W}_1$ (80×1024) $\boldsymbol{W}_2$ (80×1024)
(5)	(1)	Feature Embedding	$\boldsymbol{W}\boldsymbol{x}$ (80×1)	$\boldsymbol{W}(80 \times 1024)$
(6)	(5)	Self Predictor	*Softmax*	-
(7)	(4)	Context Predictor	*Softmax*	-

Table 1. The detailed network architecture of our VC R-CNN.

VC R-CNN的详细流程：

1. 对于任一图像经过特征提取backbone得到两个RoI特征向量x和y，对应表中的(1)(2)；
2. 对y经过SDPA得到精炼后的confounder，对应表中的(3)，对x直接做embedding用于自身类别预测，对应表中的(5)；
3. 根据线性模型$f_y(x,z)$计算X和Z对Y的因果效应，对应表中的(4)；
4. 最后将(4)和(5)输入到对应的分类器中预测各自的类别。

由于COCO数据集中涵盖了动物、植物、汽车、生活用品等80中类别，将数据集中所有的对象全部视为混杂因子的假设不符合现实场景，（例如高脚杯、人、公交车）根据D分离准则，如果在原本相互独立的两个对象之间施加控制，将会造成它们彼此相关。所以，作者在原模型的基础上进一步分析实现了Neural Causation Coefficient(NCC)模块用来消除干预对撞带来的相关性。对于给定的x和z，$NCC(x \to z)$输出x到z的相关因果强度，然后根据阈值(0.001)丢弃具有强对撞因果强度的训练样本。NCC的训练是通过两层embedding和两层classification得到一个三元分类结果(causal, anticausal, no causation)，在测试阶段就可以直接用来评价RoI特征。

> 原文是以人工合成样本来对NCC进行说明，NCC的输入样本集$S_i = \{(x_{ij}, y_{ij})\}_{j=1}^{m_i}$，其中从$Gaussian(0, r_i)$和$Gaussian(0, s_i)$中采样$k_i$个高斯分布的均值和方差，混合采样得到$x_{ij}$，再对$x_{ij}$添加异方差加性噪声得到$y_{ij}$。

基于反事实的场景图生成

> **反事实定理**
>
> 1.如果变量集Z满足$X \to Y$的后门条件，那么对于所有可能的x，反事实Y_x都与X以Z为条件独立，即：
> $P(Y_x|X,Z) = P(Y_x|Z)$
> 该定理可以直接得到一个有关估计$P(Y_x = y)$的式子：
>
> $$\begin{aligned} P(Y_x = y) &\overset{\text{全概率}}{=} \sum_z P(Y_x = y|Z = z)P(Z = z) \\ &\overset{\text{定理1}}{=} \sum_z P(Y_x = y|Z = z, X = x)P(Z = z) \\ &\overset{\text{一致性}}{=} \sum_z P(Y = y|Z = z, X = x)P(Z = z) \end{aligned} \tag{8}$$
>
> 2.定义X对Y的总效应的斜率τ如下：
>
> $$\tau = E[Y|do(x+1)] - E[Y|do(x)] \tag{9}$$

> 那么，对于任何已知的证据$Z=e$都有：
>
> $$E[Y_{X=x}|Z=e]=E[Y|Z=e]+(x-E[X|Z=e]) \tag{10}$$

Scene Graph Generation(SGG)是对图像中物体及其关系的视觉检测任务，全面的视觉场景表示可以支持图像推理的高级任务，如图像标注和视觉问答。作者认为目前的SGG仍无法实现这一承诺，因为数据标注中的bias和long-tail effect，大部分基于message passing的模型优化已经逐渐变成了更好的拟合数据集的bias，而非提取真正有意义的relationships。基于上述原因，作者结合因果关系设计了一个unbias的inference算法，尝试提取更有意义的relationship。

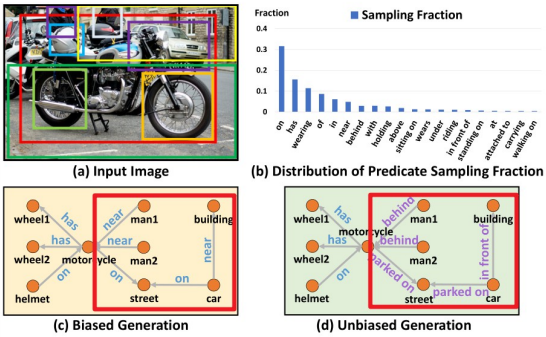


Figure 1. An example of scene graph generation (SGG). (a) An input image with bounding boxes. (b) The distribution of sample fraction for the most frequent 20 predicates in Visual Genome [22]. (c) SGG from re-implemented MOTIFS [71]. (d) SGG by the proposed unbiased prediction from the same model.

人类在自然存在的bias中出生和长大，在拥抱美好的同时避免了糟糕的context，并在content上做出无偏见的决定，这主要归功于因果关系(*causality-based*)：**决策是通过追求内容引起的主要因果关系而不是取决于上下文的副作用。**然而，及其通常是基于似然估计(*likelihood-based*)：**这种预测类似于在一个巨大的似然表中查找内容及其上下文，并通过人工训练进行插值。**在SGG任务中relationship是一个非常主观且很依赖语境的标签，而现在深度学习学出来的模型，更像是passive的只靠直觉驱使的，无法作出更加主动和主观的决策。作者认为现在大部分reasoning模型是在给定同一份输入，一个确定网络永远只会输出一个相同的结果，是一个被动的反馈。而作者通过对因果图的干预，是同一份输入，同一个网络，得到不同的输出，用于不同的目的，这样的结果更类似人对同一个问题产生各种可能性的权衡。于是，作者对现有SOTA模型进行图形化得到如下因果图，用来分析这些模型中的bias产生的原因。因果图中每个节点为一些关键变量，而其中的有向边就是对各种网络forward运算的简化，仅体现了一种因果上的决定关系。

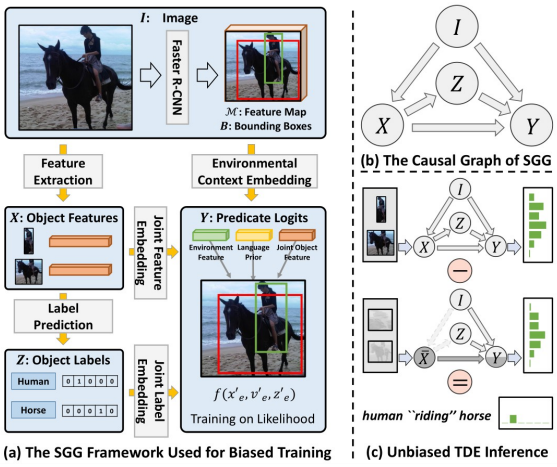


Figure 4. (a) The framework used in our biased training. (b) The causal graph of the SGG framework. (c) An illustration of the proposed TDE inference.

经过分析作者认为，SGG中的偏见主要来源于下图右下角的那种场景，即不看具体的两个物体的状态(feature)，单纯通过两个物体的label和一个union box的环境信息，就盲猜这两个物体是什么relationship。因为VisualGenome数据集的bias和长尾效应，偏偏这种盲猜不仅更容易学习还大部分情况下都是对的。这就导致了模型不再关注物体具体的状态而直接take了盲猜的biased shortcut。导致的

结果就是，具体的visual feature不再重要，也就预测不出真正有意义的fine-grain relationships了。因为更fine-grain relation出现太少，而且很容易错，所以干脆把所有复杂的sitting on/standing on/riding全预测成on。

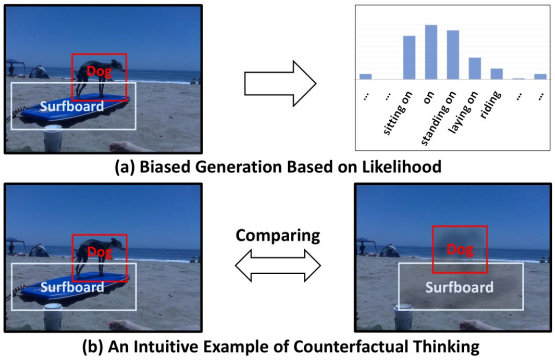


Figure 2. (a) The biased generation that directly predicts labels from likelihood. (b) An intuitive example of the proposed total direct effect, which calculates the difference between the real scene and the counterfactual one. Note that the "wipe-out" is only for the illustrative purpose but not considered as visual processing.

基于以上实验，作者引入Total Direct Effect (TDE)来取代单纯的网络log-likelihood，为模型赋予反事实因果关系的能力，以追求无偏预测中的main effect：*If I had not seen the content, would I still make the same prediction?*

传统的有偏预测模型像黑盒一样，只能看到给定整幅图像的输出结果，无法知道特定的一对对象是如何影响它们之间的relationship。然而，从因果推理角度则有助于摆脱黑盒，直接操作其中的节点，得到相应的因果效应。

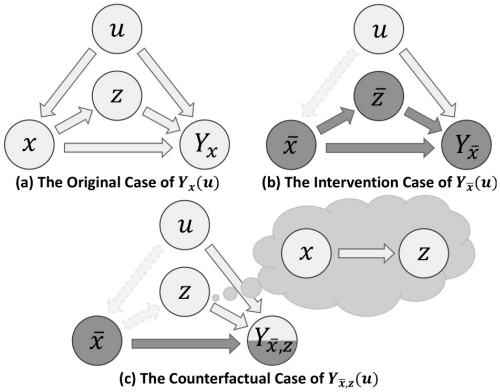


Figure 5. The original causal graph of SGG together with two interventional and counterfactual alternates.

因果推理：使用成对对象特征作X为干预的控制变量，评估其效果，对于pair不存在的将不会产生任何有效关系，干预值$\bar{x}$被设置为训练集的平均特征或是零向量，则在反事实条件下的Y表示为：

$$Y_{\bar{x}}(u) = Y(do(X = \bar{x})|u) \tag{11}$$

无偏预测不是倾向于有偏的静态似然，而是在于观察结果$Y_x(u)$和反事实$Y_{\bar{x},z}(u)$之间的差异，后者是作者希望从预测中消除的上下文特点偏差。作者提出的无偏预测是从空白到观察到具有特定特征的真实物体的视觉刺激，而不仅仅是来自周围环境和语言先验。而该想法的实现则是利用**Total Direct Effect:**

$$TDE = Y_x(u) - Y_{\bar{x},z}(u) \tag{12}$$

> **TE分解：**
>
> $$TE = Y_x(u) - Y_{\bar{x}}(u) \tag{13}$$
>
> *Decomposition 1:*

$$
\begin{aligned}
TE &= TDE + NIE \\
TDE &= Y_x(u) - Y_{\bar{x},z}(u) \\
NIE &= Y_{\bar{x},z}(u) - Y_{\bar{x}}(u)
\end{aligned} \tag{14}
$$

Decomposition 2:

$$
\begin{aligned}
TE &= TIE + NDE \\
TIE &= Y_x(u) - Y_{x,\bar{z}}(u) \\
NDE &= Y_{x,\bar{z}}(u) - Y_{\bar{x}}(u)
\end{aligned} \tag{15}
$$

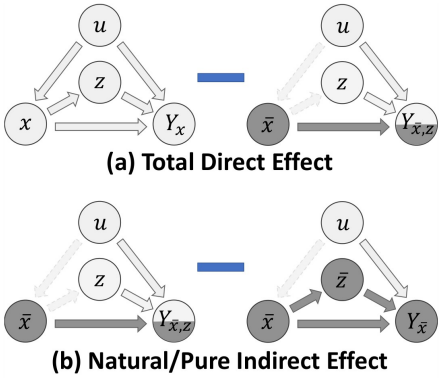


Figure 9. The illustration of Total Direct Effect and Pure/Natural Indirect Effect on causal graph.

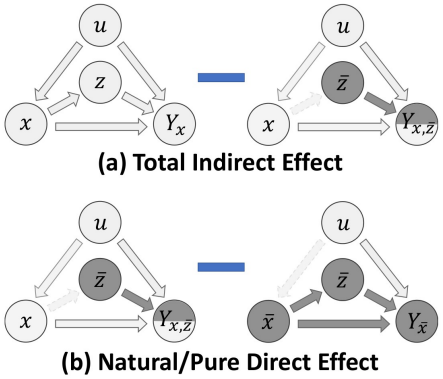


Figure 10. The illustration of Total Indirect Effect and Pure/Natural Direct Effect on causal graph.