Review: Towards Trustworthy Explanation: On Causal Rationalization

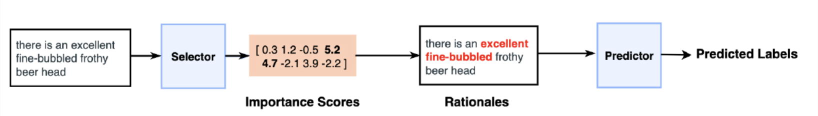
Shin Yun Seop

December 12, 2023

Seoul national university, statistics, IDEA LAB

Background: Select Rationalization

- It is a select-predict based approach which aims to find a small subset of the input that can provides similar prediction as the full input.
- It's a association based method because its main objective is selecting important features by maximize prediction accuracy

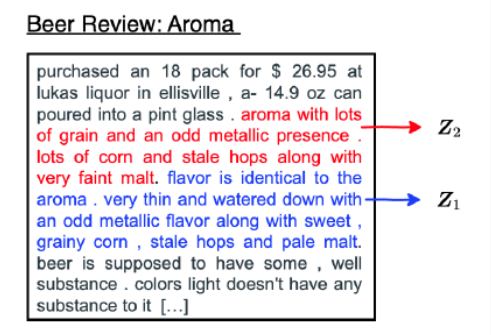


Problem: Spurious rationales

- Spurious rationales that may be related to the outcome of interest but do not indeed cause the the outcome.

Causal Interpretability is what we need!

Problem: Example

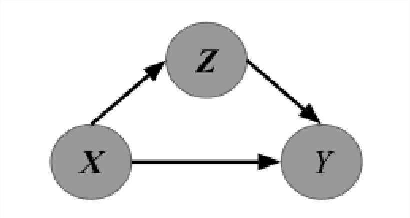


The texts of Z_1 and Z_2 are highly correlated with each other, which makes them indistinguishable in terms of predicting the sentiment of interest.

Desiderata for Rationalization

1. If we only maximize prediction accuracy, we can't avoid to select spurious rationales.
2. We now formalize the desiderate for rationalization: **non-spuriousness** and **efficiency.**
   2.1 **Non-spuriousness**: Selector can select the part of input which can causally determine the label
   2.2 **Efficient rationales**: Selector can select only essential and non-redundant part of the input.

Diagram



1. $\mathbf{X} = (X_1, \cdots, X_d)$ is input text with d tokens
2. $\mathbf{Z} = (Z_1, \cdots, Z_d)$ is corresponding selections where $Z_i \in \{0, 1\}$ indicates whether the i-th token is selected or not by the selector. And its a similar rule with treatment effect in causal inference.
3. $Y(\mathbf{Z} = \mathbf{z})$ denote the potential value of Y when setting $\mathbf{Z}$ as $\mathbf{z}$

Probability of Causation for Rationales

Conditional probability of necessity and sufficiency for single rationales defined as

$$\begin{aligned}\mathrm{CPNS}_j \triangleq P\,(\,& Y\,(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}) = y, \\ & Y\,(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}) \neq y \mid \boldsymbol{X} = \boldsymbol{x})\,.\end{aligned}$$

This can be regarded as a good proxy of causality.

Identifiability assumption

1. Consistency: $\boldsymbol{Z} = \boldsymbol{z} \rightarrow Y(\boldsymbol{Z} = \boldsymbol{z}) = Y$.
2. Ignorability:

$$\begin{aligned} &\{Y\left(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right), \\ &Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right)\} \perp \boldsymbol{Z} \mid \boldsymbol{X}. \end{aligned}$$

Calculate the CPNS

Theorem

Assume the causal diagram in page 5 holds.

1. *If assumptions 1 and 2 hold,* CPNS_j *is not identifiable but its lower bound can be calculated by*

$$\underline{\mathrm{CPNS}}_j = \max\left[0, P\left(Y = y \mid Z_j = z_j, \boldsymbol{Z}_{-j} = z_{-j}, \boldsymbol{X} = \boldsymbol{x}\right) - P\left(Y = y \mid Z_j \neq z_j, \boldsymbol{Z}_{-j} = z_{-j}, \boldsymbol{X} = \boldsymbol{x}\right)\right].$$

Objective function

$$\min_{\theta,\phi} \mathcal{L} = \min_{\theta,\phi} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[L(y,\widehat{y}) + \lambda\delta(\boldsymbol{z}) \underbrace{- \mu \sum_{j\in\mathbf{r}^{(k)}} \frac{\log \widehat{\mathrm{CPNS}}_j^+}{\left|\mathbf{r}^{(k)}\right|}}_{\text{Causality Constraint}}],$$

where $\widehat{y} = h_\phi(\boldsymbol{z} \odot \boldsymbol{x}), \boldsymbol{z} = g_\theta(\boldsymbol{x}), L(\cdot,\cdot)$ defined as the cross-entropy loss.

Objective function

1. $\delta(\cdot)$ is the sparsity penalty to control sparseness of rationales.
2. $\mathbf{r}_i^{(k)}$ denotes the a random subset with size equal $k\%$ of the sequence length.
   The reason we sample a random subset $\mathbf{r}_i^{(k)}$ is due to the computational cost of flipping each selected rationale.
3. λ and μ are the tuning parameters.
4. $\underline{\widehat{\mathrm{CPNS}}}_j^{+} = \underline{\widehat{\mathrm{CPNS}}}_j + 1$

Algorithm i

algorithm | Causal Rationalization

Require: Training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, parameters

Begin: Initialize the parameters of selector $g_\theta(\cdot)$ and predictor $h_\phi(\cdot)$, where θ and ϕ denote their parameters

while not converge **do**

Sample a batch $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ from $\mathcal{D}$

Generate selections $\mathcal{S} = \{\boldsymbol{z}_i\}_{i=1}^{n}$ through Gumbel- Softmax sampling

Algorithm ii

for $i = 1, \cdots n$ **do**

Get a random sample $\mathbf{r}_i^{(k)}$ from index set $\mathbf{r}_i$ where $\mathbf{r}_i$ represents the set of rationales that are selected as 1 in z_i and its size equals $k\% \times$ length $(\mathbf{x}_i)$

for $j = 1, \cdots \left|\mathbf{r}_i^{(k)}\right|$ **do**

Generate counterfactual selections $\boldsymbol{z}_{i(j)}$ by flip- ping the j th index of the index set $\mathbf{r}_i^{(k)}$

end for

end for

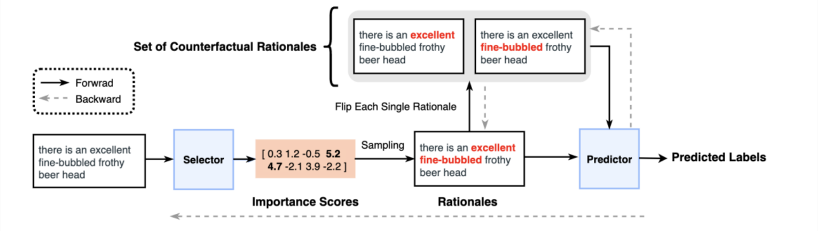
Get a new batch of selections $\tilde{\mathcal{S}} = \left\{\boldsymbol{z}_{i(j)}\right\}_{j=1,\cdots,\left|\mathbf{r}_i^{(k)}\right|}^{i=1,\cdots,n}$ and set $\mathcal{S}_{\text{all}} = \mathcal{S} \cup \tilde{\mathcal{S}}$ Compute $\mathcal{L}$ via page(9) by using $\mathcal{S}_{\text{all}}$ and $\mathcal{D}$

Update parameters $\theta = \theta - \alpha \nabla_\theta \mathcal{L}; \phi = \phi - \alpha \nabla_\phi \mathcal{L}$

end while

Output: selector $g_\theta(\cdot)$ and predictor $h_\phi(\cdot)$

Framework of algorithm



Suggestions for Future Research

What if assumption does not hold??
$\Rightarrow$ Sensitivity analysis is possible?

Reference

1. Zhang, Wenbo, Tong Wu, Yunlong Wang, Yong Cai, and Hengrui Cai. "Towards Trustworthy Explanation: On Causal Rationalization." ArXiv.org (2023): ArXiv.org, 2023. Web.