Inducing Causal Structure for Interpretable Neural Networks (ICML 2022)

SeongSik Choi

January 2, 2024

Seoul National University

Notation

$\mathcal{V}$: a set of variables

For variable $V \in \mathcal{V}$, $\mathrm{Val}(V)$: a set of values

PA_V : a set of parents

F_V : a structural equation that sets the value of V based on the setting of its parents.

V_{In} : the set of variables with no parents

V_{Out} : those with no children.

A structural causal model $\mathcal{M} = (\mathcal{V}, PA, \mathrm{Val}, F)$ can represent both symbolic computations and neural networks.

Notation(GetVals)

We define GetVals($\mathcal{M}$, inp, V) $\in$ Val(V) : the particular values that V takes on when the model $\mathcal{M}$ processes inp.

For example, $\mathcal{M}$ could correspond to structure and weight parameters in a neural network and V could correspond to nodes in a neural network.

For a set of variables V and a setting for those variables $v \in$ Val(V), we define $\mathcal{M}_{V \leftarrow v}$ to be the causal model identical to $\mathcal{M}$, except that the structural equations for V are set to constant values v.

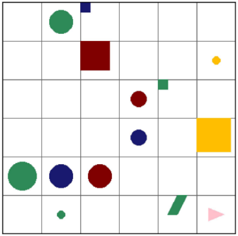
Notation(Interchange Intervention)

$$\text{INTINV}(\mathcal{M}, \text{base}, \text{source}, \mathbf{V}) \stackrel{\text{def}}{=}$$
$$\text{GetVals}\big(\mathcal{M}_{\text{V} \leftarrow \text{GetVals}(\mathcal{M}, \text{ source, V})}, \text{base}, \mathbf{V}_{\text{Out}}\big)$$

In short, the interchange intervention provides the output of the model $\mathcal{M}$ for the input base, except the variables **V** are set to the values they would have if source were the input.

Example : Navigation and Language (ReaSCAN)

The goal is to predict an action sequence for the agent to reach the referred target and operate on it given a command and a grid world.



I_{world}(World) : Figure

I_{com}(Command) : "walk the cylinder"

O(Action sequence) : 'turn left', 'turn left', 'walk'

Training

For each variable V in $\mathcal{C}_{\text{ReaSCAN}}$ aligned with neurons $\mathbf{N}_V$ in $\mathcal{N}^{\theta}_{\text{CNN-LSTM}}$, we optimize for $\mathcal{N}^{\theta}_{\text{CNN-LSTM}}$ implementing the marginalized submodel $\mathcal{C}^{V}_{\text{ReaSCAN}}$:

$$\sum_{\text{b,s} \in \text{ReaSCAN}} \text{CE}_{\text{Action}} \left(\text{INTINV}\left(\mathcal{N}^{\theta}_{\text{CNN-LSTM}}, \mathbf{b}, \mathbf{s}, \mathbf{N}_V\right), \right.$$
$$\left. \text{INTINV}\left(\mathcal{C}^{V}_{\text{Rea-SCAN}}, \mathbf{b}, \mathbf{s}, V\right) \right),$$

where $\text{CE}_{\text{Action}}$ is the cross-entropy loss over each action token prediction over the complete action sequence.

Correspondence between V and N_V

The correspondence between variable V and node N_V is as shown in the following diagram.