Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness

Raphael Suter [1], Đorđe Miladinović [1], Bernhard Schölkopf[2], Stefan Bauer [2]
[1]ETH Zurich, [2]MPI for Intelligent Systems

ICML 2019

Contributions

- Causal Model for Representation Learning
- Interventional Robustness Score
- Visualising Robustness

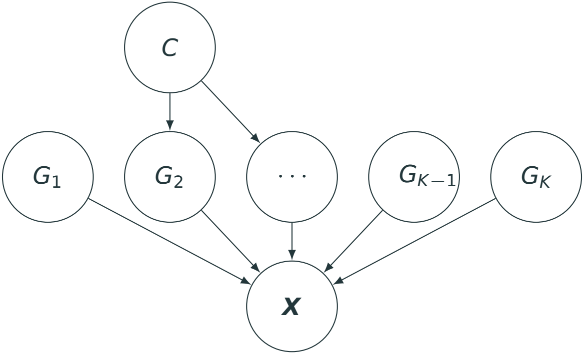
Disentangled Representations

Observation: $\boldsymbol{X} \in \mathbb{R}^n$

Feature encoding: $\boldsymbol{Z} = E(\boldsymbol{X}) \in \mathbb{R}^K, n \gg K$

Disentanglement $\iff$ components Z_i represent different sources of variation in $\boldsymbol{X}$

Definition: Disentangled Causal Process

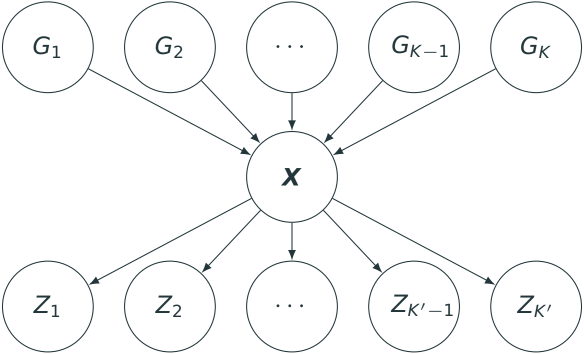


Disentangled Causal Mechanisms:

$$\forall g_j^{\triangle} \quad p(g_i|\mathrm{do}(G_j \leftarrow g_j^{\triangle})) = p(g_i) \quad \left(\neq p(g_i|g_j^{\triangle})\right)$$

Unified Causal Model

Generative Factors

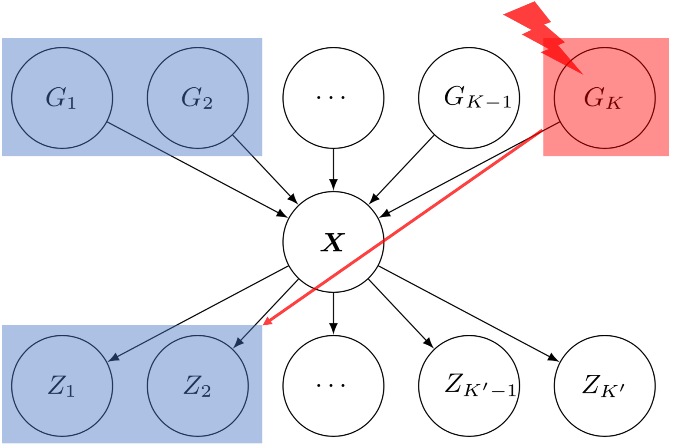


Feature Representation

Robust Representation

relevant factors: G_1, G_2

nuisance factor: G_K



selected features: Z_1, Z_2

Interventional Robustness

Post Interventional Disagreement

$$d\left(\mathbb{E}[\boldsymbol{Z}_{sel}|\boldsymbol{g}_{rel})], \mathbb{E}[\boldsymbol{Z}_{sel}|\boldsymbol{g}_{rel}, \mathrm{do}(\boldsymbol{G}_{nuis} \leftarrow \boldsymbol{g}^{\triangle}_{nuis})]\right)$$

Interventional Robustness Score

normalised score $\in [0, 1]$

Theoretical Results

- Properties of a disentangled causal process
- IRS estimation from observational data $\mathcal{D} = \{(\boldsymbol{g}^{(i)}, \boldsymbol{x}^{(i)})\}_{i=1}^{N}$
- Handles confounding $G_i \leftarrow C \rightarrow G_j$
- Efficient $\mathcal{O}(N)$ algorithm

Conclusion

- disentanglement_lib by Locatello et al. (2019): `github.com/google-research/disentanglement_lib`
- Poster: Thurs 06:30 – 09:00 PM at Pacific Ballroom #29