Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness

Raphael Suter\textsuperscript{1}, Đorđe Miladinović\textsuperscript{1}, Bernhard Schölkopf\textsuperscript{2}, Stefan Bauer\textsuperscript{2}  
\textsuperscript{1}ETH Zurich, \textsuperscript{2}MPI for Intelligent Systems

ICML 2019
Contributions

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

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

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

**Disentanglement** $\iff$ components $Z_i$ represent different sources of variation in $\mathbf{X}$
Definition: Disentangled Causal Process

Disentangled Causal Mechanisms:

$$\forall g_j \triangledown \quad p(g_i | \text{do}(G_j \leftarrow g_j \triangledown)) = p(g_i) \quad (\neq p(g_i | g_j \triangledown))$$
Unified Causal Model

Generative Factors

$G_1 \quad G_2 \quad \cdots \quad G_{K-1} \quad G_K$

Feature Representation

$Z_1 \quad Z_2 \quad \cdots \quad Z_{K'-1} \quad Z_{K'}$
Robust Representation

relevant factors: $G_1, G_2$

nuisance factor: $G_K$

selected features: $Z_1, Z_2$
Interventional Robustness

Post Interventional Disagreement

\[ d \left( \mathbb{E}[Z_{sel}|g_{rel}], \mathbb{E}[Z_{sel}|g_{rel}, \text{do}(G_{nuis} \leftarrow g_{nuis})] \right) \]

Interventional Robustness Score

normalised score \( \in [0, 1] \)
Theoretical Results

• Properties of a disentangled causal process
• IRS estimation from observational data
  \[ D = \{(g^{(i)}, x^{(i)})\}_{i=1}^N \]
• Handles confounding \( G_i \leftarrow C \rightarrow G_j \)
• Efficient \( \mathcal{O}(N) \) algorithm
• disentanglement_lib by Locatello et al. (2019):
  github.com/google-research/disentanglement_lib
• Poster: Thurs 06:30 – 09:00 PM at Pacific Ballroom #29