Proceedings of Machine Learning Research TBD:1–3, 2025 4th Conference on Causal Learning and Reasoning

CLEAR-VAE: Causal Learning and Explanation with Attributed Representations

Author Name1 ABC@SAMPLE.COM
Address 1

Author Name2 XYZ@SAMPLE.COM
Address 2

Editors: Biwei Huang and Mathias Drton

Theorem 1 (Concept Identifiability) *Let $\mathbf{z} = [\mathbf{z}_y, \mathbf{z}_d]$ represent the latent decomposition into domain-invariant $\mathbf{z}_y$ and domain-specific $\mathbf{z}_d$ components. Assume encoder functions $q_{\phi_y}(\mathbf{z}_y|\mathbf{x})$ and $q_{\phi_d}(\mathbf{z}_d|\mathbf{x})$ and decoder function $decoder(\cdot)$. If the following conditions hold:*

1. ***Orthogonality of Concept Representations**: The concept separation loss satisfies:*

$$\mathcal{L}_{concept} = 1 - \cos\left(normalize(\mathbf{c}_y), normalize(\mathbf{c}_d)\right) \leq \varepsilon_1,$$

ensuring that domain-invariant $\mathbf{c}_y$ and domain-specific $\mathbf{c}_d$ representations are orthogonal. This constraint is inspired by prior work [Von Kügelgen et al., 2021] that uses data augmentations to achieve invariance. Here, we replace augmentation-based constraints with explicit orthogonality in the latent space, extending the identifiability guarantees.

2. ***Consistency with Prior Distributions**: The KL divergence terms for the latent spaces are bounded:*

$$D_{KL}(q_{\phi_y}(\mathbf{z}_y|\mathbf{x}) \parallel p(\mathbf{z}_y)) \leq \varepsilon_2 \quad \textit{and} \quad D_{KL}(q_{\phi_d}(\mathbf{z}_d|\mathbf{x}) \parallel p(\mathbf{z}_d)) \leq \varepsilon_2,$$

ensuring that the latent distributions are consistent with the assumed priors, following the principles of β-VAE [Higgins et al., 2017].

3. ***Reconstruction Fidelity**: The reconstruction loss satisfies:*

$$\|\mathbf{x} - decoder(\mathbf{z}_y, \mathbf{z}_d)\|^2 \leq \varepsilon_3,$$

ensuring that the learned latent representations preserve sufficient information to accurately reconstruct the input data.

Under these conditions, and assuming a well-specified model and sufficient data, the learned concepts $\mathbf{c}_y$ and $\mathbf{c}_d$ are identifiable up to isometry, aligning with the identifiability results in causal representation learning [Daunhawer et al., 2023].

Theorem 2 (Convergence Guarantees) *Let the total loss function be defined as:*

$$\mathcal{L}_{total} = \lambda_r \mathcal{L}_{recon} + \mathcal{L}_{KL} + \lambda_y \mathcal{L}_{digit} + \lambda_d \mathcal{L}_{domain} + \lambda_c \mathcal{L}_{concept},$$

where $\mathcal{L}_{recon}$ is the reconstruction loss, $\mathcal{L}_{KL}$ is the KL divergence regularizer, and the remaining terms correspond to task-specific and interpretability losses. The learning rates η_t satisfy:

$$\eta_t \to 0, \quad \sum_{t=1}^{\infty} \eta_t = \infty, \quad \textit{and} \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty.$$

If the weights $\lambda_r, \lambda_y, \lambda_d, \lambda_c$ are appropriately balanced, the optimization of $\mathcal{L}_{total}$ converges to a local minimum with rate $\mathcal{O}(1/T)$. This result builds upon the convergence guarantees for variational models in [Kingma et al., 2015] and extends them to disentangled representations incorporating orthogonal constraints.

0.1. Connection to Prior Work and Novel Interpretability

While our work builds upon the theoretical foundations established in ”Self-supervised learning with data augmentations provably isolates content from style”, we extend this framework in several important directions focusing on interpretability and explainability:

Theorem 3 (Connection to Style-Content Separation) *Given the style-content separation framework from prior work that establishes:*

$$\mathbb{E}_{\tau \in \mathcal{T}}[\|h(x) - h(\tau(x))\|] \leq \epsilon \tag{1}$$

where h is the content encoder and τ represents augmentations, our LLM-guided approach enhances this by:

1) Introducing explicit concept interpretability:

$$\mathcal{L}_{concept} = \mathbb{E}_{x,y}[KL(p_{LLM}(c|x,y) \| q_\phi(c|h(x)))] \tag{2}$$

where p_{LLM} represents LLM-guided concept descriptions and q_ϕ is our learned concept distribution.

2) Enforcing semantic alignment through LLM guidance:

$$sim(c_{inv}, LLM(y)) \geq \alpha \text{ and } sim(c_{var}, LLM(d)) \geq \beta \tag{3}$$

for some thresholds $\alpha, \beta > 0$, where c_{inv} and c_{var} are our invariant and variant concepts.

This formulation provides several key advantages:

1. **Interpretable Disentanglement**: While prior work focused on statistical independence, our approach enforces semantic interpretability through LLM guidance.

2. **Concept Grounding**: The LLM provides natural language grounding for both invariant (digit-specific) and variant (rotation-specific) concepts, making the learned representations more accessible to human understanding.

3. **Verifiable Separation**: Through our concept probing mechanism, we can verify that the learned representations align with human-interpretable concepts, going beyond purely statistical measures of disentanglement.

Proposition 4 (Interpretability Guarantee) *Given LLM-guided concept extraction $\mathcal{L}_{LLM}$ and domain separation loss $\mathcal{L}_{sep}$, our model guarantees:*

1) Semantic Consistency: $\mathbb{P}(LLM(decode(c_{inv})) \approx LLM(y)) \geq 1 - \delta$

2) Style-Content Separation: $I(c_{inv}; d|y) \leq \epsilon$ *and* $I(c_{var}; y|d) \leq \epsilon$

where $\delta, \epsilon > 0$ *are small constants.*

This theoretical framework extends prior work by not only ensuring statistical disentanglement but also providing guarantees about human interpretability and semantic meaning in the learned representations.

References