Debiasing Concept-based Explanations with Causal Analysis

Taha Bahadori & David Heckerman

May 3, 2021

Amazon

Incorporating Domain Knowledge via Concepts

Features ⟶ **Concepts** ⟶ **Labels**

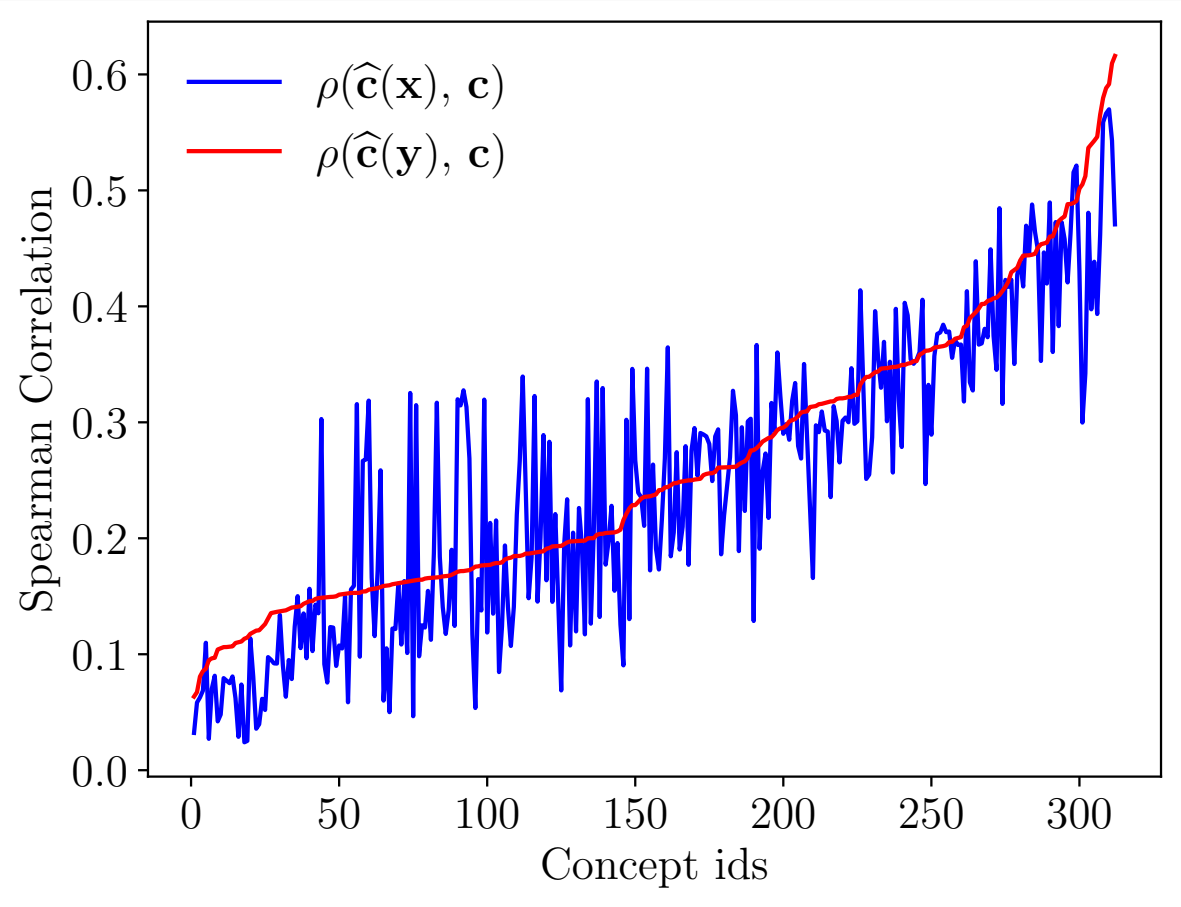
has throat color::black
has forehead color::black
has crown color::black
has breast color::red
has belly color::white
has underparts color::white
has underparts color::red
has back color::black
has primary color::white
has bill shape::cone
has breast pattern::multi-colored
has primary color::red

Rose Breasted Grosbeak

- Concepts allows analysis of models on complex manifolds.
- Concepts bring domain knowledge into the explanation process.

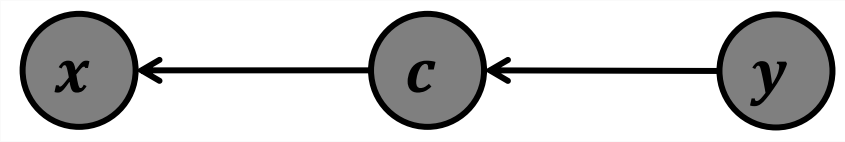
Confounding and Noise Bias in Concept-Based Explanations

Spearman correlation coefficients (ρ) of the predictors of the concepts given features $\widehat{\mathbf{c}}(\mathbf{x})$ and labels $\widehat{\mathbf{c}}(\mathbf{y})$.



$\widehat{\mathbf{c}}(\mathbf{x})$ captures spurious correlation and isn't just predictive of $\mathbf{y}$.

Causal Prior Graph for the Case without Confounding



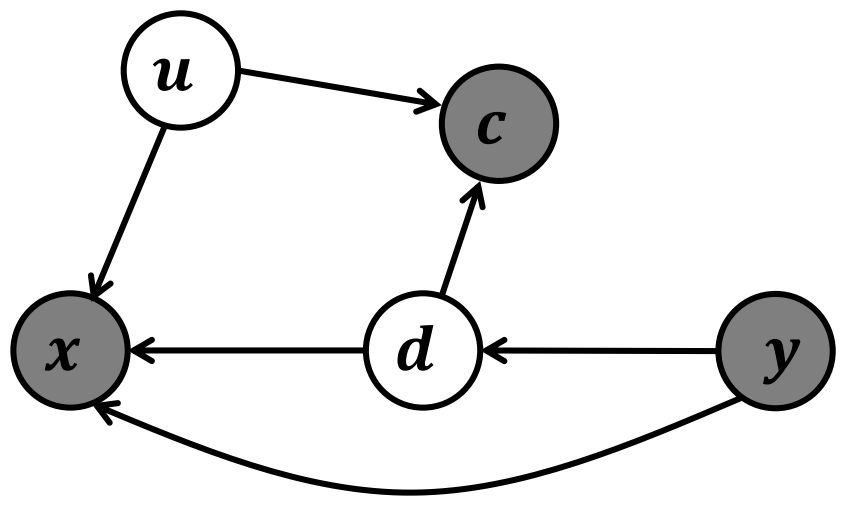
Generative model:

- Generate labels $\mathbf{y}$ randomly.
- Generate concepts $\mathbf{c} \sim p(\mathbf{c}|\mathbf{y})$.
- Generate images from concepts $\mathbf{x} \sim p(\mathbf{x}|\mathbf{c})$.

Drawbacks:

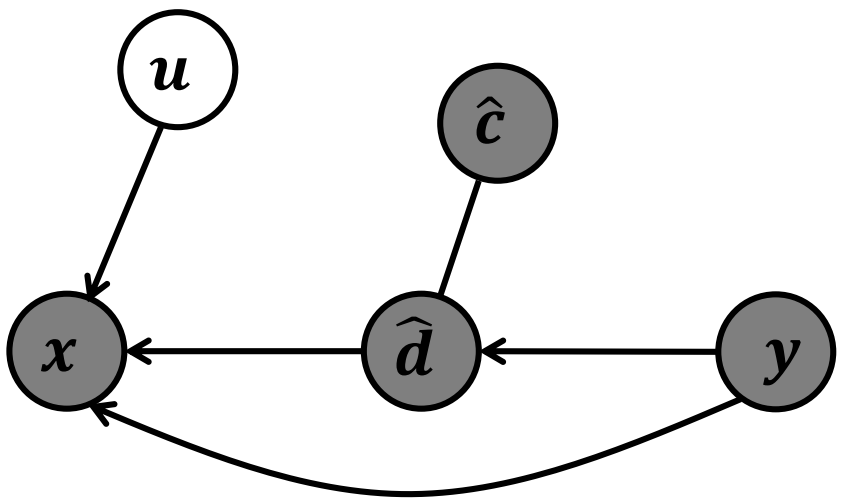
- No shared context information between $\mathbf{x}$ and $\mathbf{c}$.
- Concept completeness

A More Realistic Causal Prior Graph



- Latent variable **u** represents the shared context between **x** and **c**.
- Discriminative concept vector **d**
- Direct **x** $\leftarrow$ **y** to capture the residual correlation between **x** and **y**

A Technique from Instrumental Variables



- An estimate for the discriminative concepts $\widehat{\mathsf{d}} = E[\mathsf{c}|\mathsf{y}]$.
- $\mathsf{u} \perp\!\!\!\perp \mathsf{y} \Longrightarrow \mathsf{u} \perp\!\!\!\perp \widehat{\mathsf{d}}$

Incorporating Concept Prediction Uncertainty

Our prediction of concepts $\widehat{\mathsf{c}}(\mathsf{x})$ can be uncertain. Incorporate uncertainty into $\widehat{\mathsf{y}}(\mathsf{x}) = \widehat{\mathsf{y}}(\widehat{\mathsf{c}}(\mathsf{x}))$.

To incorporate the uncertainty in our estimation:

$$E[\mathsf{y}|\mathsf{x}] = E[g_{\boldsymbol{\theta}}(\widehat{\mathsf{d}})|\mathsf{x}] = \int g_{\boldsymbol{\theta}}(\boldsymbol{d})\mathrm{d}p_{\boldsymbol{\phi}}(\mathsf{d} = \boldsymbol{d}|\mathsf{x}),$$

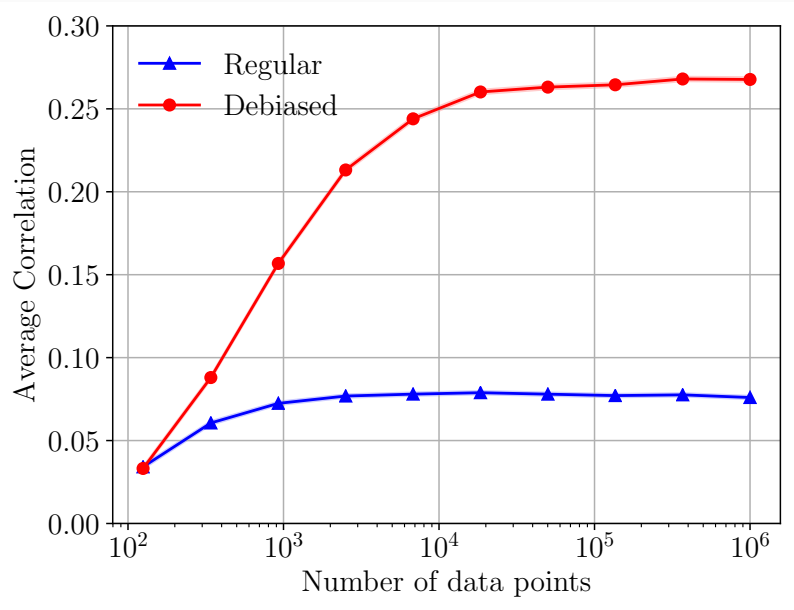
The integral is computed using Monte Carlo method.

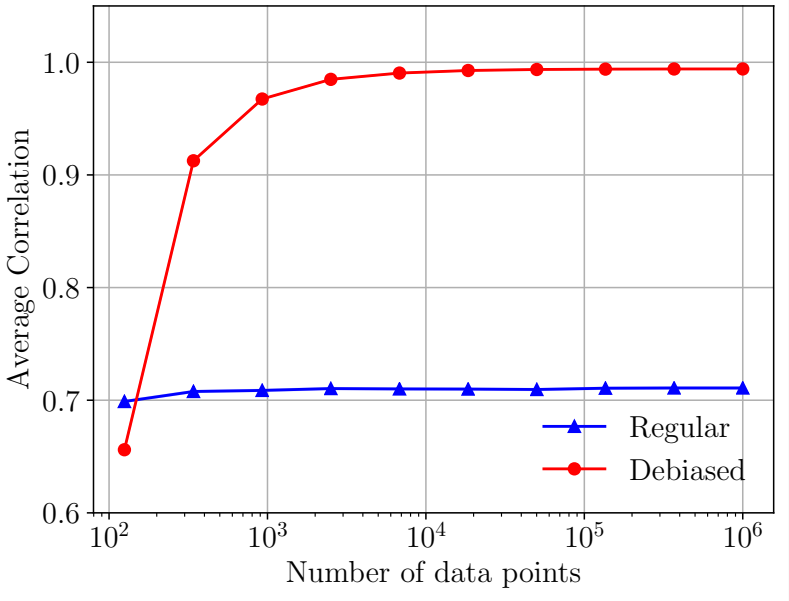
Synthetic Data

Our causal prior graph with linear transformations:

- Generate n vector pairs $\boldsymbol{y}_i, \boldsymbol{u}_i \in \mathbb{R}^{100}$ with elements $\sim \mathcal{N}(0, 1)$.
- Generate n noise vector pairs $\boldsymbol{\varepsilon}_{c,i}, \boldsymbol{\varepsilon}_{x,i} \in \mathbb{R}^{100}$ with elements $\sim \mathcal{N}(0, \sigma = 0.02)$.
- Generate matrices $W_{\mathbf{y}\to\mathbf{d}}, W_{\mathbf{u}\to\mathbf{c}}, W_{\mathbf{d}\to\mathbf{x}}, W_{\mathbf{u}\to\mathbf{x}} \in \mathbb{R}^{100\times100}$ with elements $\sim \mathcal{N}(0, \sigma = 0.1)$.
- Compute $\boldsymbol{d}_i = W_{\mathbf{y}\to\mathbf{d}}\boldsymbol{y}_i + \boldsymbol{\varepsilon}_{d,i}$ for $i = 1, \ldots, n$.
- Compute $\boldsymbol{c}_i = \boldsymbol{d}_i + W_{\mathbf{u}\to\mathbf{c}}\boldsymbol{u}_i$ for $i = 1, \ldots, n$.
- Compute $\boldsymbol{x}_i = W_{\mathbf{d}\to\mathbf{x}}\boldsymbol{d}_i + W_{\mathbf{u}\to\mathbf{x}}\boldsymbol{u}_i + \boldsymbol{\varepsilon}_{x,i}$ for $i = 1, \ldots, n$.

Results





Regular design

orthogonal ($W_{y \to d} \perp W_{u \to c}, W_{u \to x}$)

CUB-200-2011 Data

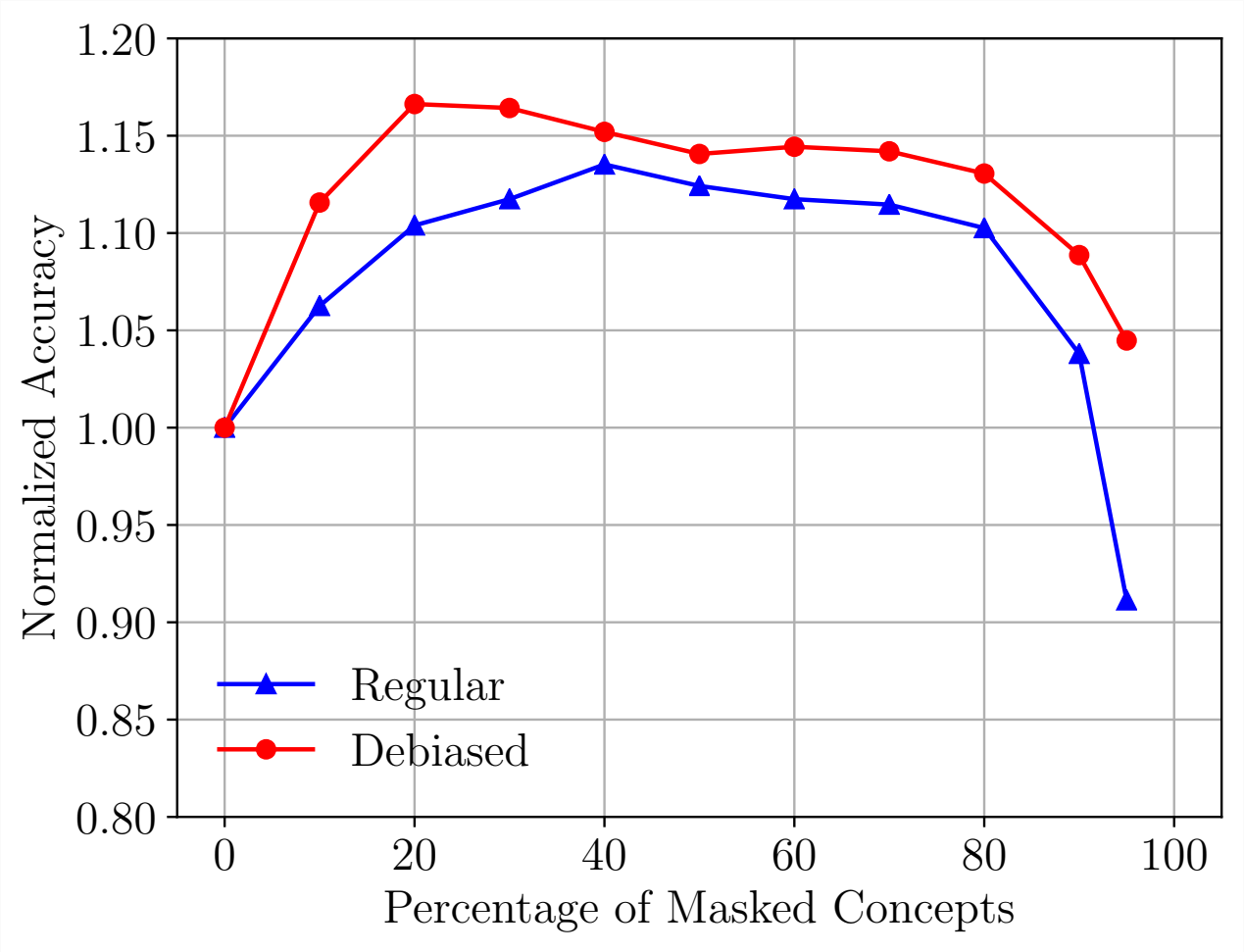
- 11788 pictures (in 5994/5794 train/test partitions)
- 200 different types of birds
- Annotations for each picture: bird type and 312 different concepts.
- Randomly choose 15% of the training set as the validation set.

Examples of concepts:

- has_bill_shape::dagger, has_bill_shape::needle
- has_wing_color::purple, has_wing_color::blue
- has_breast_pattern::solid, has_breast_pattern::spotted

ROAR Evaluation

Top-5 accuracy of label prediction improved from 39.5% to 49.3%.



ROAR: RemOve And Retrain

- Mask bottom x% of concepts and retrain the $\mathbf{c} \rightarrow \mathbf{y}$ predictor.