IMPERIAL

LAIDLAW SCHOLARS

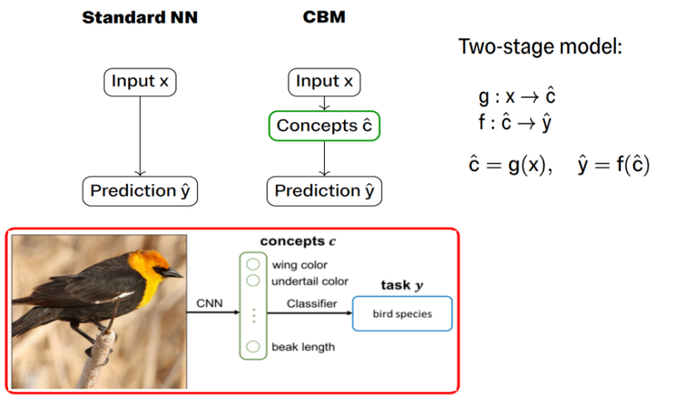
Beyond the Black Box: Making AI Clinically Trustworthy Using Group-Aware Concept Bottleneck Machines

Author: Pranav Sharma
Supervisor: Dr Sonali Parbhoo
AI for Actionable Impact Lab, Imperial College London

Introduction

An AI tells a doctor a patient is high-risk. The obvious next question is: why?

Modern machine-learning models can make remarkably accurate predictions but often provide little visibility into the reasoning behind them. In high-stakes healthcare settings, that creates a fundamental problem: a prediction cannot be safely acted on if the person responsible for the decision cannot understand, challenge or correct it.

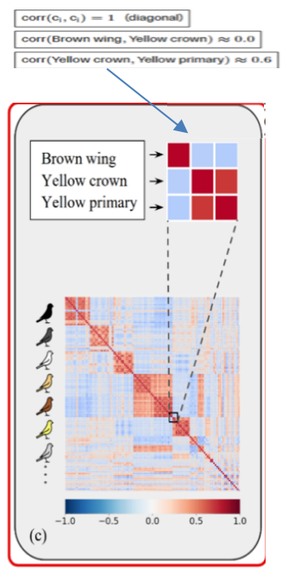


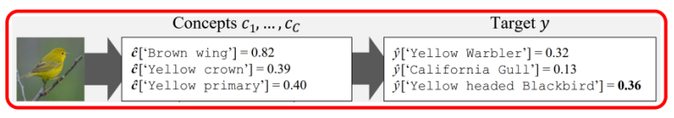
Concept Bottleneck Models (CBMs) address this by routing prediction through an **intermediate layer of human-interpretable concepts**, that a clinician can inspect and, when necessary, correct.

Stochastics CBMs replace hard concept labels with uncertain, probabilistic predictions.
Each concept is represented **not as a fixed value, but as a Gaussian distribution:**

$$\hat{c} \sim \mathcal{N}(\mu(x), \Sigma)$$

capturing both uncertainty in each concept and correlations between them. A covariance matrix Σ encodes how concepts relate.





Motivation

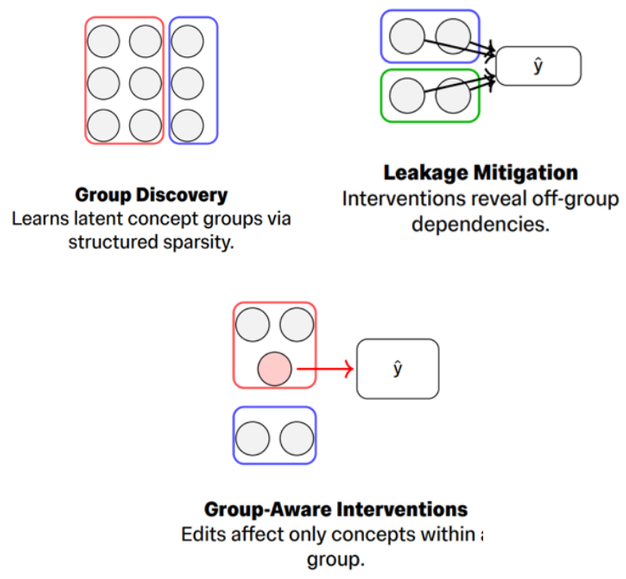
The three limitations of CBMs and SCBMs:

- Concept Dependency Assumption*:* CBMs assume concepts independent to one another while SCBMs assume global concept dependence
- Concept Leakage: Model learns to exploit patterns as label hints, bypassing the intended interpretable reasoning.
- Concept Completeness: Even perfect scores can't help if crucial concepts are missing.

Thus we build a **structured model of dependency**, one that captures local correlations without enforcing global connectivity: Group-Aware Stochastic Concept Bottleneck Machines.

Aims

Grouped Stochastic Concept Bottleneck Model (GSCBM)



Methodology

The GSCBM is designed to address key limitations of traditional and stochastic concept bottleneck models by introducing structure, modularity, and diagnostic capacity into the concept space. This is achieved through three stages:

Stage 1: Structured Sparsity with Known Concept Groups

- Proof of concept for incorporating structured sparsity into precision matrix leveraging known concept groupings.
- Provide the model with ground-truth groups assignments.

Stage 2: Soft Group Discovery with Fixed Group Count

- Remove the assumption of known group assignments while still specifying the number of groups K.
- Precise groupings are unknown and must be discovered from data.

Stage 3: Full Bayesian Group Discovery with Hyperprior on Group Count

- Remove the assumption of a fixed number of groups. Model learns both the concept-to-group assignments and the effective number of groups directly from data, reflecting real-world scenarios.

Evaluation Metrics

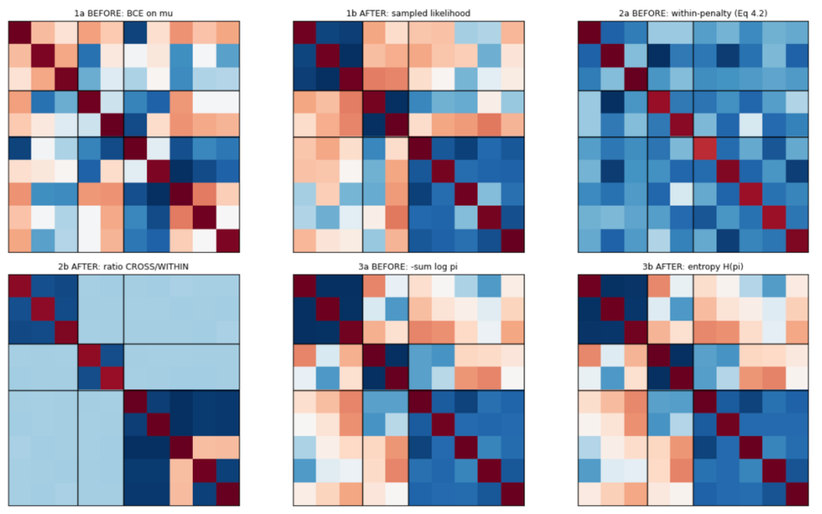
To avoid evaluating an unsupervised algorithm using supervised metrics, I used cross group mutual Information and modularity how much more internally connected groups are than chance would predict.

$$I(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

In order to evaluate the unsupervised nature of Stage 3, we use a hyperprior on number of groups & groups memberships using the Chinese Restaurant Process to evaluate outputs.

$$(a)_{m,k} = \prod_{i=0}^{m-1}(a+ik) = \begin{cases} a^m & \text{if } k = 0, \\ k^m \left(\frac{a}{k}\right)^{\overline{m}} & \text{if } k > 0, \\ |k|^m \left(\frac{a}{|k|}\right)^{\underline{m}} & \text{if } k < 0 \end{cases}$$

Results



tag	off_frob	within	cross	MI_true_groups	Q_true_groups	D_true_groups	concept	target
1a BEFORE: BCE on mu	0.7723	4.775	3.717	0.090	−0.034	−44.928001	0.880	0.825
1b AFTER: sampled likelihood	0.4450	13.797	4.072	0.355	0.234	20.201000	0.912	0.800
2a BEFORE: within-penalty (Eq 4.2)	0.8953	0.110	0.137	0.070	−0.094	−2.121000	0.920	0.925
2b AFTER: ratio CROSS/WITHIN	0.0050	5.706	0.017	0.000	0.371	38.015999	0.922	0.750
3a BEFORE: −sum log pi	0.4336	13.078	3.470	0.249	0.248	32.848999	0.920	0.775
3b AFTER: entropy H(pi)	0.4316	13.161	3.472	0.249	0.249	33.250000	0.920	0.775

Stage 1's **core hypothesis confirmed**, using a loss function ratio that encourages within group linkage and penalises cross group linkage.

Given true membership, block structure is almost fully recovered (**off-block connection error: 0.005**). **92% of concept accuracy achieved.**

Stage 3's internal hyperprior **shifted to entropy giving a peaked distribution** rather than an even, indecisive one.

Conclusion and Future Work

We have validated the core hypothesis that Group Aware Concept Bottleneck Machines can address current limitations and recover structured sparsity given ground truth groups. Going forward we look to:

1. Empirically test the model to recover the number of groups and group memberships given no ground truth
2. Run synthetic tests, where group membership is ambiguous and compare performance across stages, run a larger number of tests establishing reliability

I'd like to express my gratitude to Dr Sonali Parbhoo for being an incredible mentor and teacher to me, to Nikita Narayanan for being an infallible source of help and to Shreya Kamath, upon whose work much of my work was built upon.