Avoiding Leakage Poisoning: Concept Interventions Under Distribution Shifts

Paper

Code

Mateo Espinosa Zarlenga[1], Gabriele Dominici[2], Pietro Barbiero[3], Zohreh Shams[1,4], Mateja Jamnik[1]

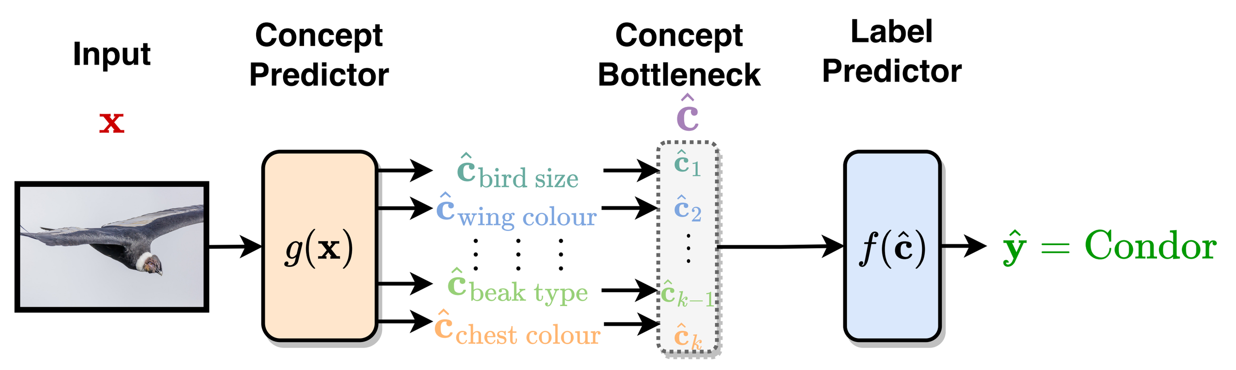
[1] University of Cambridge, [2] Università della Svizzera italiana, [3] IBM Research, [4] Leap Laboratories Inc.

ICML International Conference On Machine Learning 2025

1 Background: CBMs

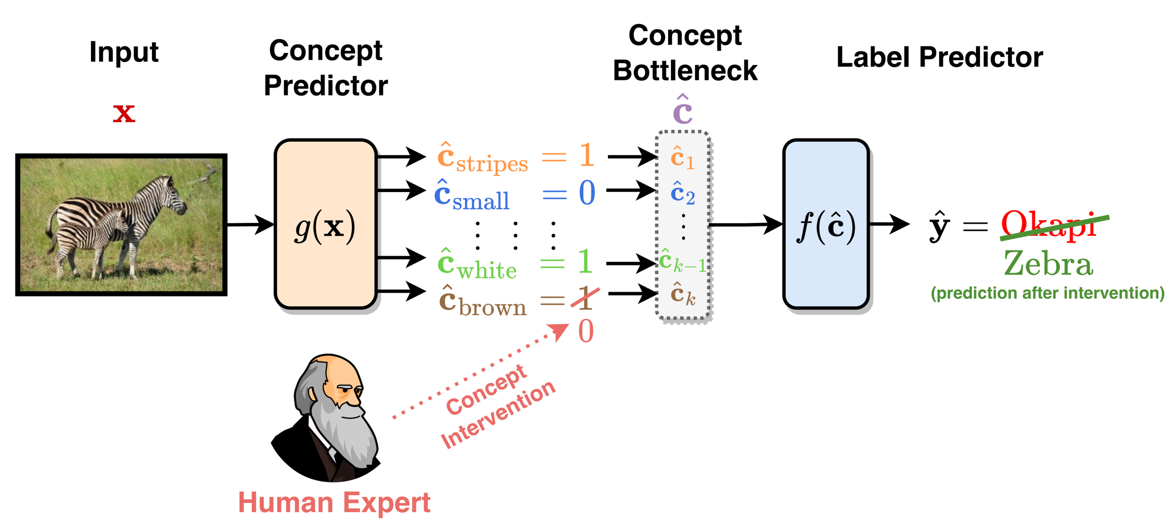
Concept Bottleneck Models (CBMs)

Concept Bottleneck Models (CBMs) [koh et al.] are a family of *interpretable* deep neural networks that, given an input $\boldsymbol{x}$, first predict a set of high-level *"concept"* representations $\hat{\boldsymbol{c}}$ and then predict a task label $\hat{y}$ from $\hat{\boldsymbol{c}}$.

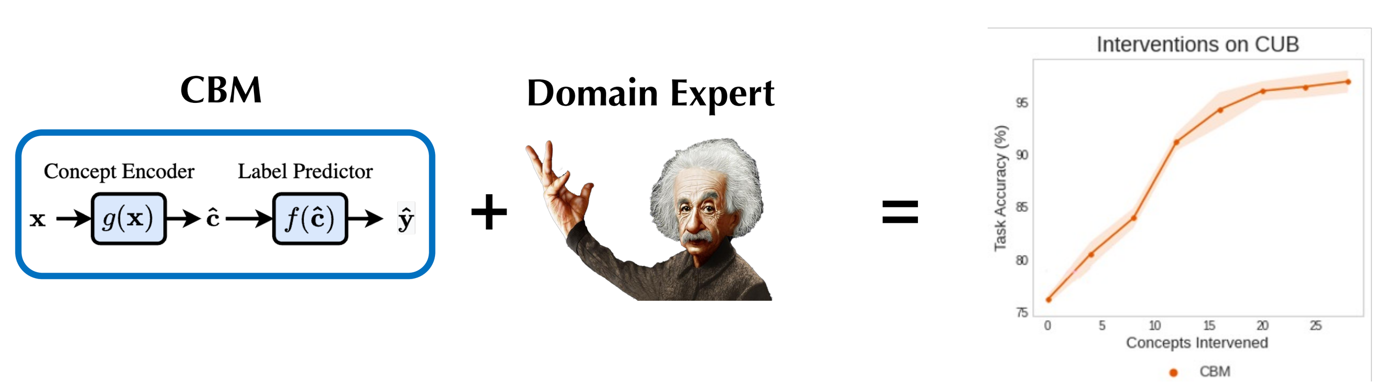


Test-time Concept Interventions

CBMs can improve their accuracy by allowing an expert to *correct mispredicted concepts at test-time:*



This can lead to powerful *collaborative* human-AI systems that can outperform the original model or expert:



2 Problem: Leakage Poisoning

Importance of Intervenability

In theory, intervenability allows CBMs to receive help for "tricky" inputs, such as **Out-Of-Distribution (OOD)** inputs

A noisy but understandable X-ray scan

A chicken in the unusual act of flying

New Tradeoffs in Intervenability

However, we show that state-of-the-art CBMs struggle to remain both *intervenable for OOD* inputs **and** accurate when their training concept set is *incomplete*.

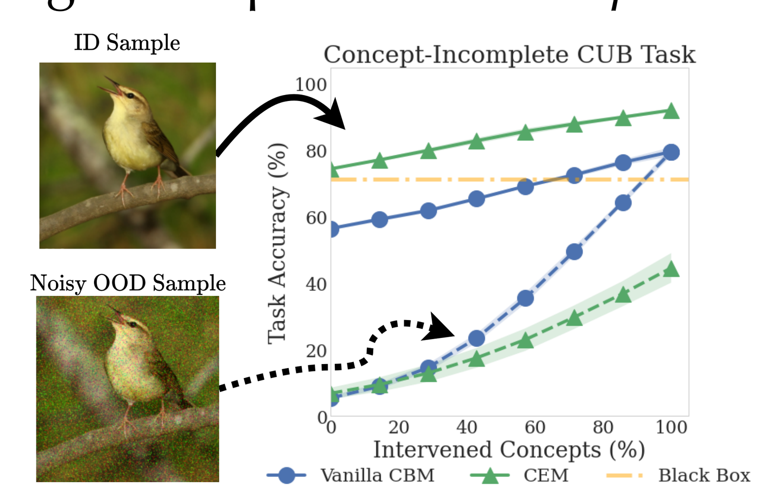
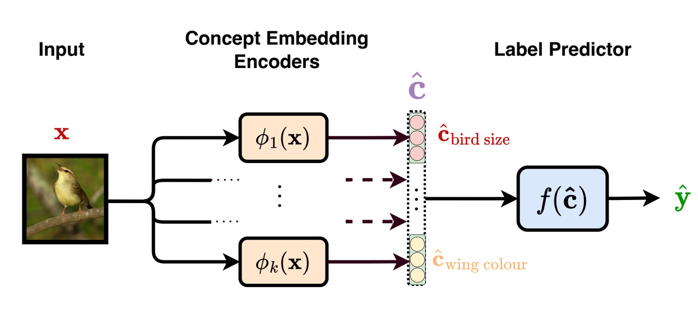
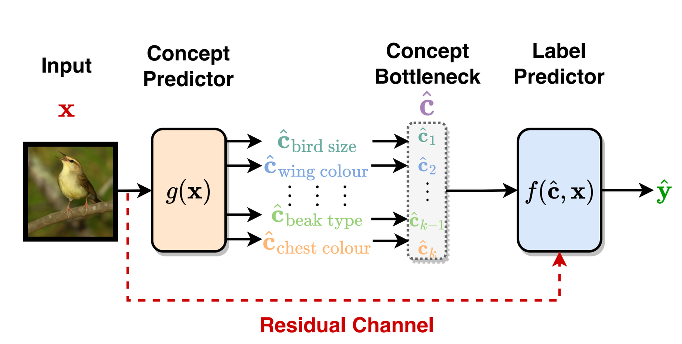


Fig 1: Task accuracy of "Vanilla CBM" and CEM, a state-of-the-art CBM, when intervening on in-distribution (solid lines) and out-of-distribution (dashed lines) test sets.

This is because existing *"completeness-agnostic"* CBMs use *information bypasses (e.g., embeddings or residuals)* that can get corrupted, or **poisonous**, for OOD inputs.



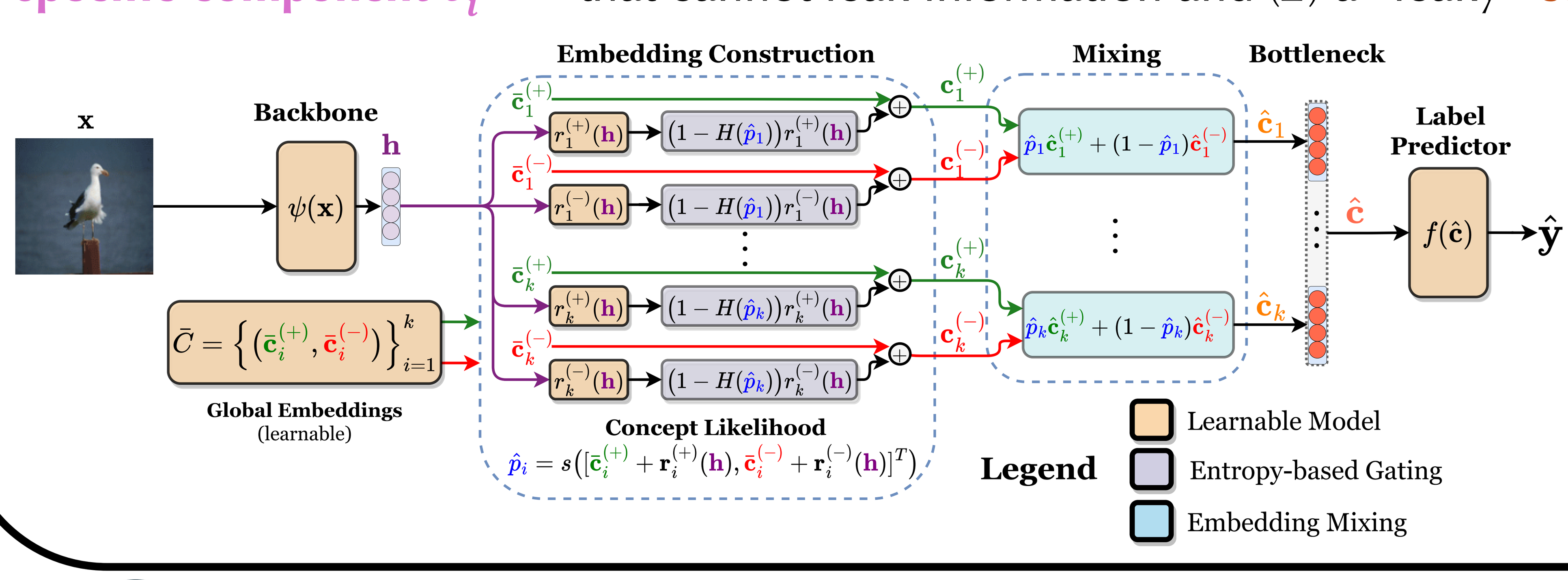
Embedding-based Models



Residual-based Models

3 Solution: Mixture of Concept Embeddings Model (MixCEM)

Our model, MixCEM, determines when leakage is *helpful* and when it is *poisonous*. It does this by learning two *embeddings* ($\boldsymbol{c}_i^{(+)}, \boldsymbol{c}_i^{(-)}$) per concept (one for when the concept is **on** and one when it is **off**) that are formed by mixing (1) a **global concept-specific component** $\bar{c}_i^{(+/-)}$ that cannot leak information and (2) a "leaky" **contextual sample-specific component** $\boldsymbol{r}_i^{(+/-)}(\boldsymbol{x})$.



$$\boldsymbol{c}_i^{(+/-)} = \bar{c}_i^{(+/-)} + \left(1 - H(\hat{p}_i)\right) r_i^{(+/-)}(\boldsymbol{x})$$

Concept embedding — Global concept-specific embedding (non-leaky) — Leakage gating mechanism (entropy-based) — Contextual sample-specific embedding (leaky)

4 Key Results

Key Result #1: MixCEM remains highly accurate and intervenable for ID and OOD test sets. This holds across different forms of distribution shifts.

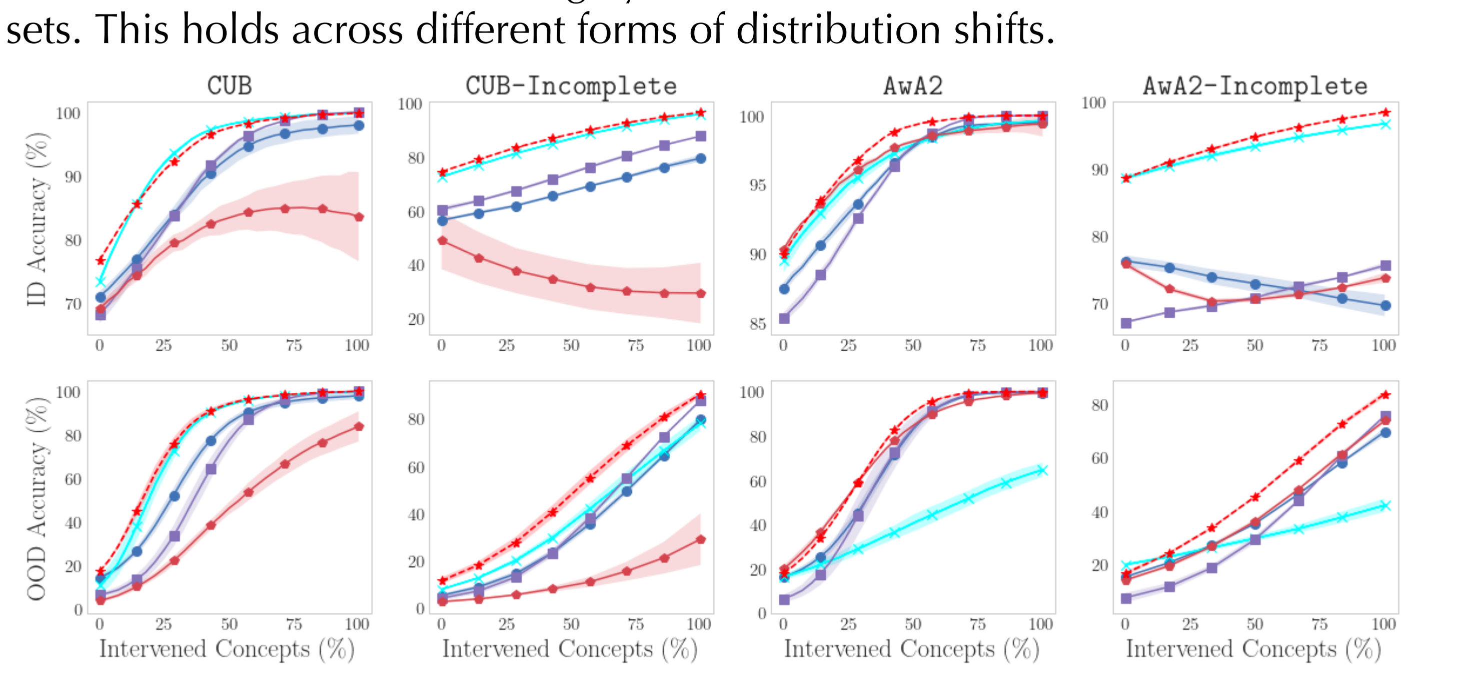


Fig 2: Intervention curves for in-distribution (top) and out-of-distribution (noised, bottom) test sets. See paper for similar results with more datasets, baselines, and forms of OOD shifts.

Key Result #2: MixCEM's bottlenecks remain in-distribution after being intervened on even for OOD samples.

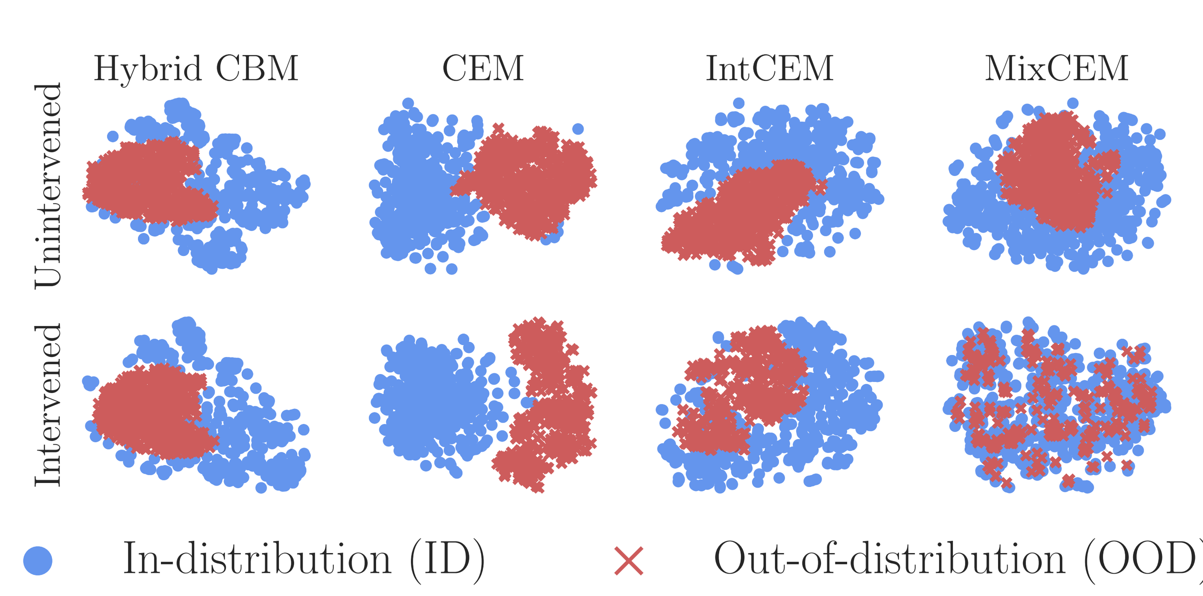


Fig 3: t-SNE projections of different concept bottlenecks before and after all concepts have been intervened on for an OOD test set.