Concept Embedding Models

Mateo Espinosa Zarlenga*, **Pietro Barbiero***, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michaelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, Mateja Jamnik

UNIVERSITY OF CAMBRIDGE
The Alan Turing Institute
GATES Cambridge
UNIVERSITÉ CÔTE D'AZUR
Inria
KATHOLIEKE UNIVERSITEIT LEUVEN
babylon
UNIVERSITÀ DI SIENA 1240
TAILOR
HUMANE AI NET
AI4media

NEURAL INFORMATION PROCESSING SYSTEMS 2022

Overview

- Concept bottleneck models (CBMs) [1] increase trustworthiness and model transparency by conditioning tasks on high-level units of information, or concepts (e.g. "whiskers").
- Nevertheless, these models **struggle to find optimal compromises** between high performance, robust explanations, and effective concept interventions in scenarios where concept annotations are scarce.
- We propose **Concept Embedding Models (CEMs)**, a novel family of CBMs that go beyond the current accuracy-vs-interpretability trade-off by learning interpretable high-dimensional concept representations.

When Concepts Annotations Are Not Enough

Concept incompleteness forces sacrifices in performance for CBMs.

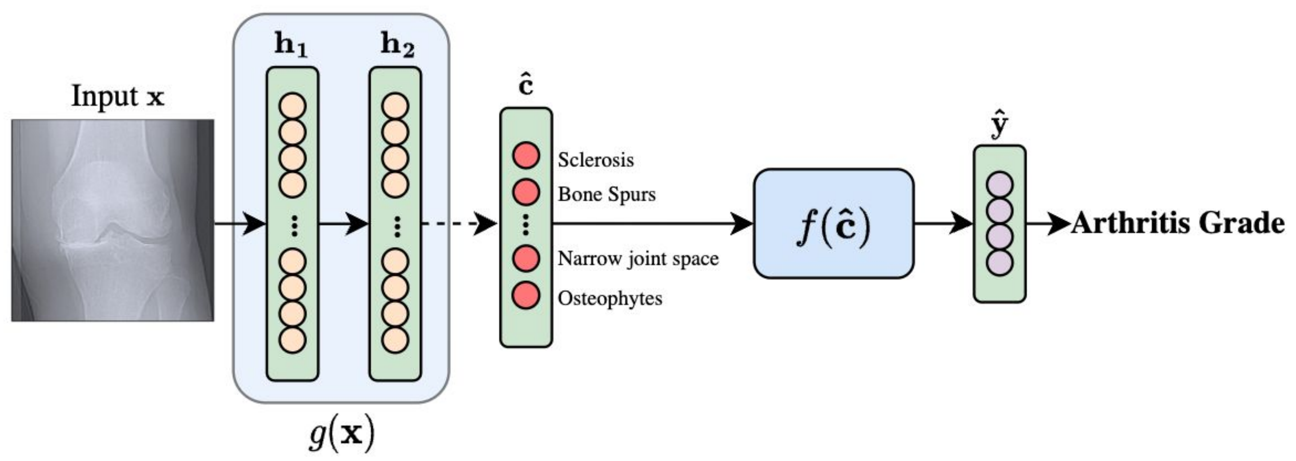

Figure 1: Concept Bottleneck Model Architecture.

- CBMs (see Figure 1) constrain their information flow so that one first predicts a set of concepts $\hat{\mathbf{c}}$ and then predicts an output label $\hat{\mathbf{y}}$ given $\hat{\mathbf{c}}$.
- This leads to a **significant practical problem**: if the set of training concepts is incomplete w.r.t. the downstream task, then the CBM is **forced to compromise either its interpretability or accuracy** (see Figure 2.a).
- On the other hand, if we naively extend the bottleneck of a CBM to allow for extra unsupervised activations (we call this a **Hybrid-CBM**), the effect of supervisions is completely lost (see Figure 2.b).

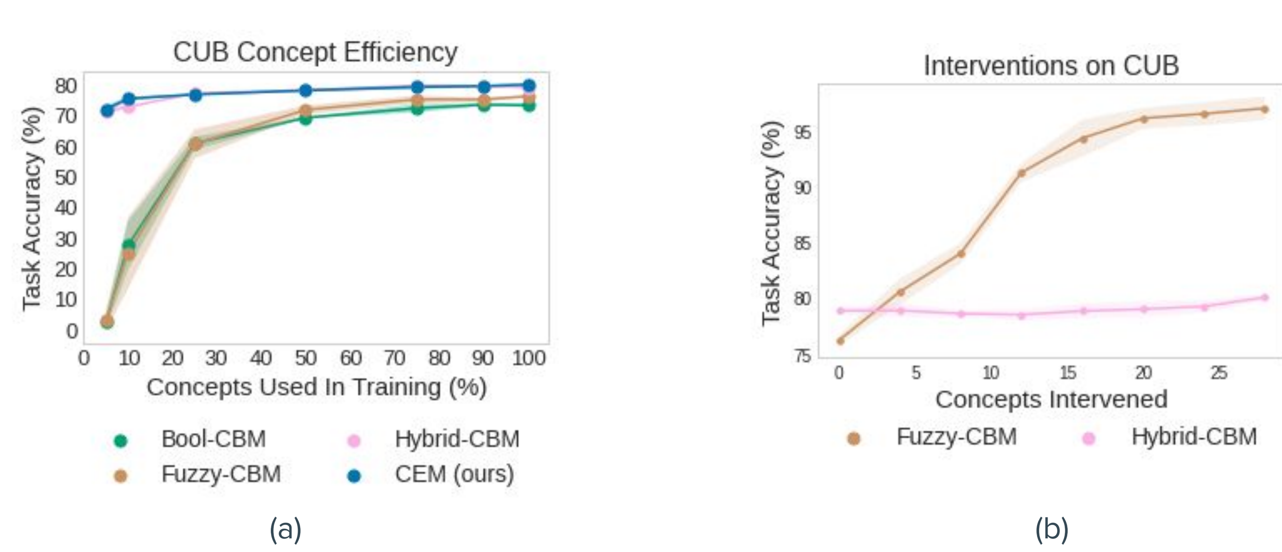

Figure 2: **(a)** Task accuracy vs percent of ground truth concepts given during training in CUB and **(b)** task accuracy vs number of concepts intervened for CBM vs Hybrid-CBM in CUB.

The Concept Embedding Model Architecture

We enable unseen concepts to be learnt in a high-dimensional embedding space while allowing for effective concept interventions.

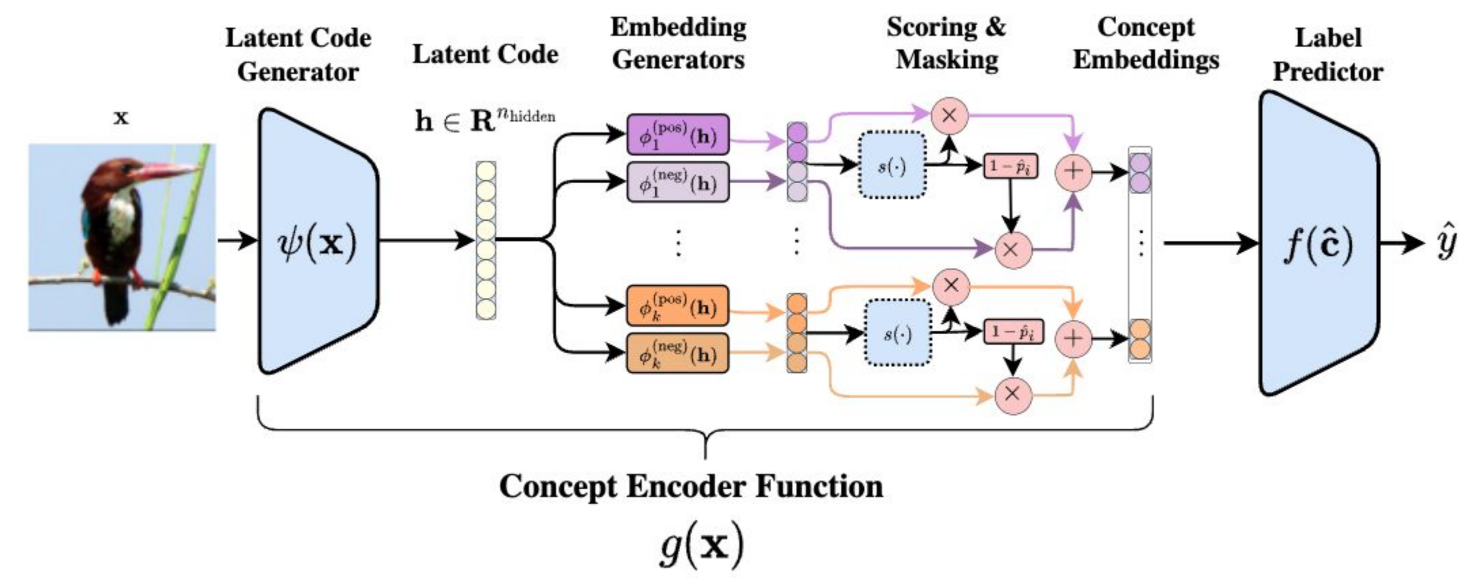

Figure 3: Concept Embedding Model (CEM) Architecture.

For each **ground-truth concept** c_i a Concept Embedding Model (CEM) learns:

1. A probability $\hat{p}_i$ that c_i is active → Can be used to **explain the end prediction.**
2. A **"negative" embedding** $\hat{\mathbf{c}}_i^-$ representing concept c_i when it is inactive.
3. A **"positive" embedding** $\hat{\mathbf{c}}_i^+$ representing concept c_i when it is active.
4. A final embedding $\hat{\mathbf{c}}_i$ representing concept c_i as a **mixture** of its positive and negative embeddings weighted by its probability of activation:

$$\hat{\mathbf{c}}_i \triangleq \left(\hat{p}_i \hat{\mathbf{c}}_i^+ + (1-\hat{p}_i)\hat{\mathbf{c}}_i^-\right)$$

Having two semantic embeddings per concept enables an **effective intervention mechanism** for CEMs while allowing the **information related to unseen concepts to flow through the bottleneck**.

Finally, **we encourage CEM to be receptive to test-time interventions** by using *RandInt*, a regularizer that randomly intervenes on CEMs bottleneck at test time:

$$\hat{\mathbf{c}}_i = \begin{cases} \left(c_i \hat{\mathbf{c}}_i^+ + (1-c_i)\hat{\mathbf{c}}_i^-\right) & \text{with probability } p_{\text{int}} \\ \left(\hat{p}_i \hat{\mathbf{c}}_i^+ + (1-\hat{p}_i)\hat{\mathbf{c}}_i^-\right) & \text{with probability } (1-p_{\text{int}}) \end{cases}$$

Avoiding the accuracy-interpretability-tradeoff

*CEM is **as accurate (or better) than black-box models** while remaining **highly interpretable**.*

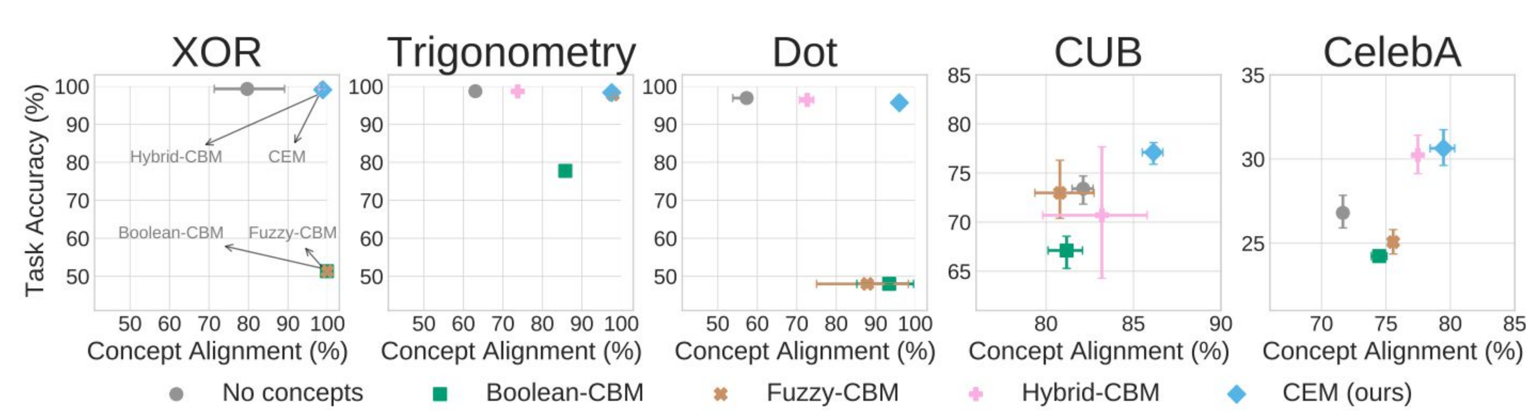

Figure 4: Task Accuracy vs Concept Alignment across all datasets and baselines.

Effective Test-time Interventions

*A CEM's **performance significantly improves via test-time concept interventions** while being more resilient to "incorrect" concept interventions.*

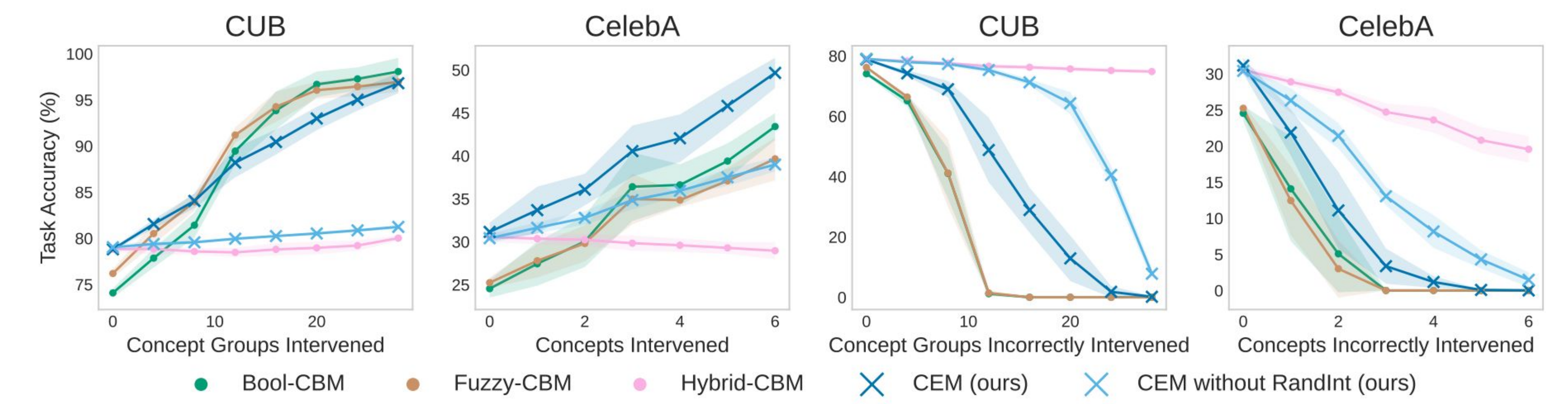

Figure 5: Task accuracy when intervening on a varying number of concepts in a model's bottleneck.

Coherent Concept Embedding Spaces

CEMs learn embeddings that are coherent w.r.t. class and concept labels.

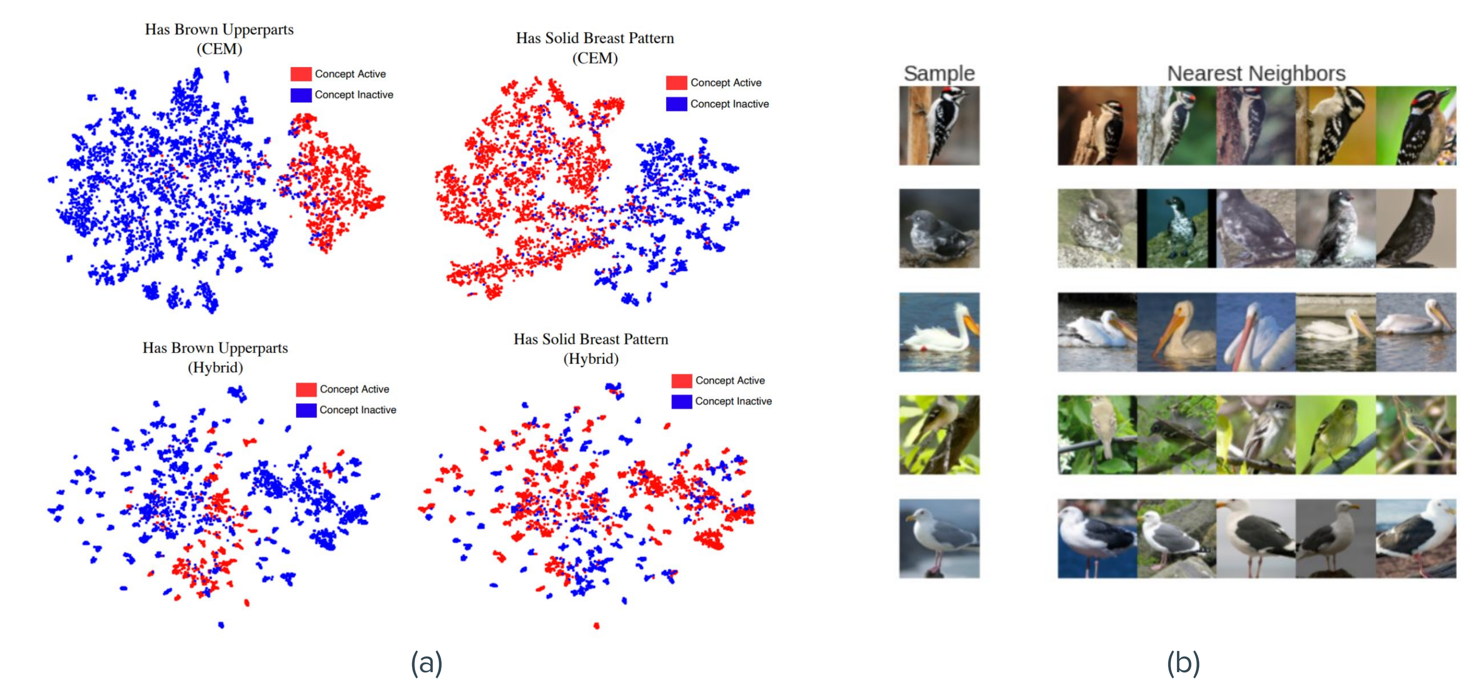

Figure 6: **(a)** T-SNE visualisation of CEM's and Hybrid-CBM's concept embedding spaces for two CUB concepts and **(b)** top-5 test neighbours of concept "has white wings" across 5 random test samples.

Ask us about...

- Why are CEMs more resilient to incorrect concept interventions?
- How does CEM's architecture affect the training dynamics through the lense of the information bottleneck?
- How difficult is it to fine-tune and train CEMs vs other CBMs?
- What opportunities does this research open?

References

[1] Koh, Pang Wei, et al. "Concept bottleneck models." In: *International Conference on Machine Learning*. PMLR, 2020.

Code + Paper