POSITION: IN DEFENSE OF INFORMATION LEAKAGE IN CONCEPT-BASED MODELS

Mateo Espinosa Zarlenga

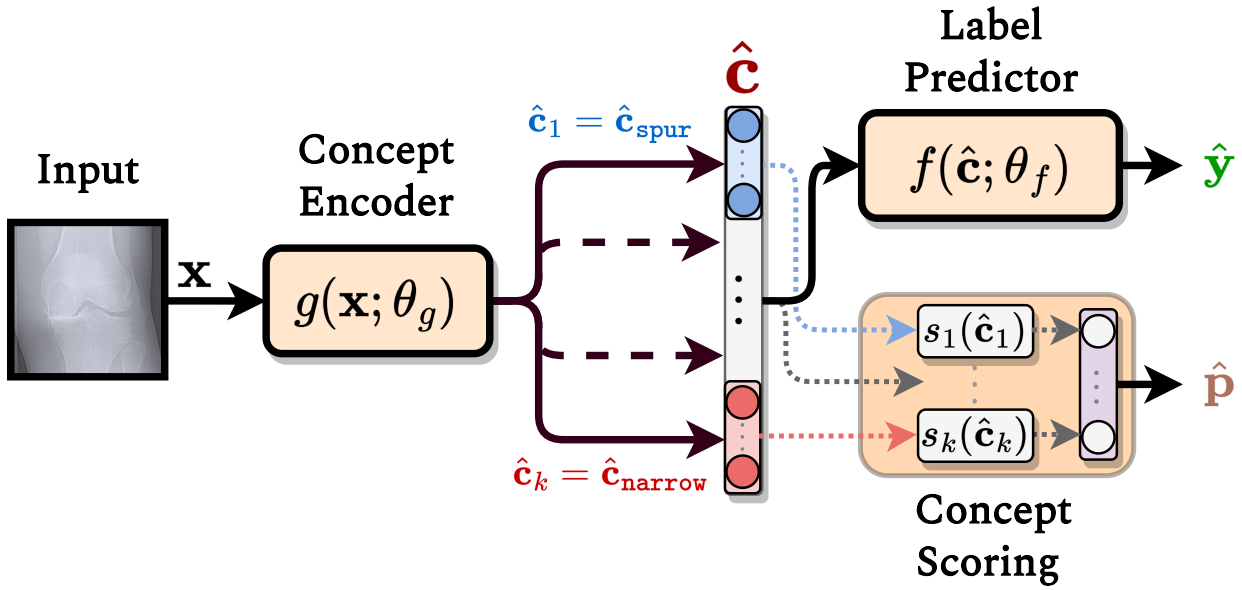
mateo.espinosazarlenga@trinity.ox.ac.uk

ICML 2026

Tu, Jul 7, 14:30 (Poster Session 4)

CONCEPT-BASED MODELS

We study models that can be framed as **Concept-based Models (CMs):**

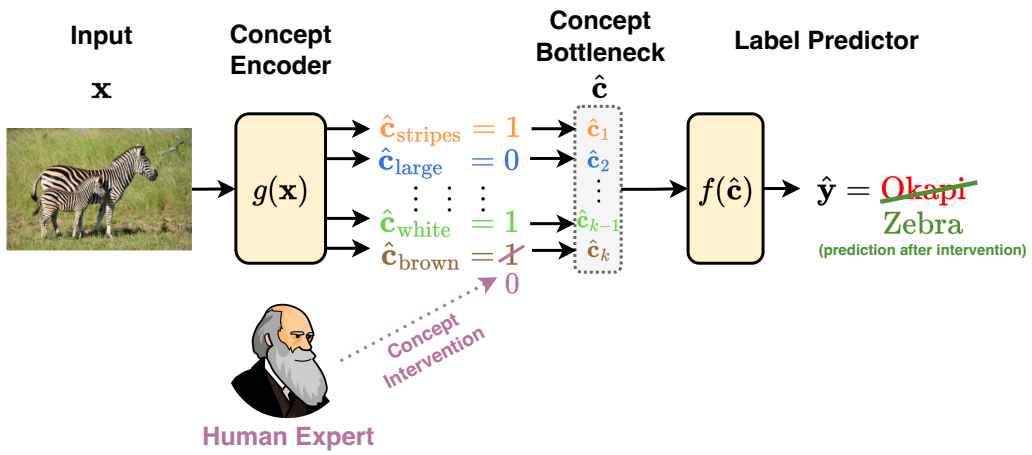


CMs are DNNs that, given an input X, ground their predictions Y on a set of representations $\hat{C}$ that are aligned with human-understandable concepts C

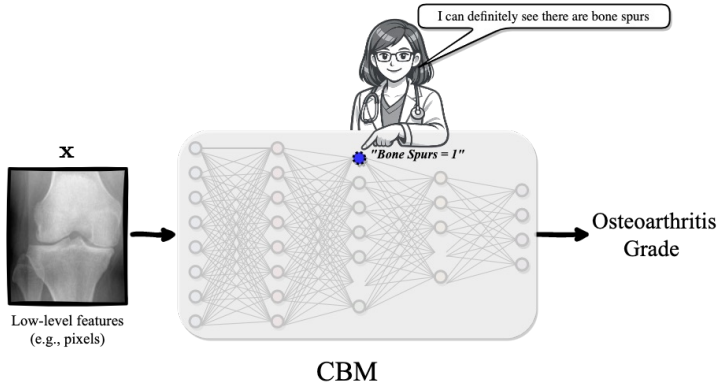
CONCEPT INTERVENTIONS

Besides explaining a prediction in terms of concepts, CMs can be intervened on their concept representations at test time to:

1 **Correct Mispredicted Concepts**

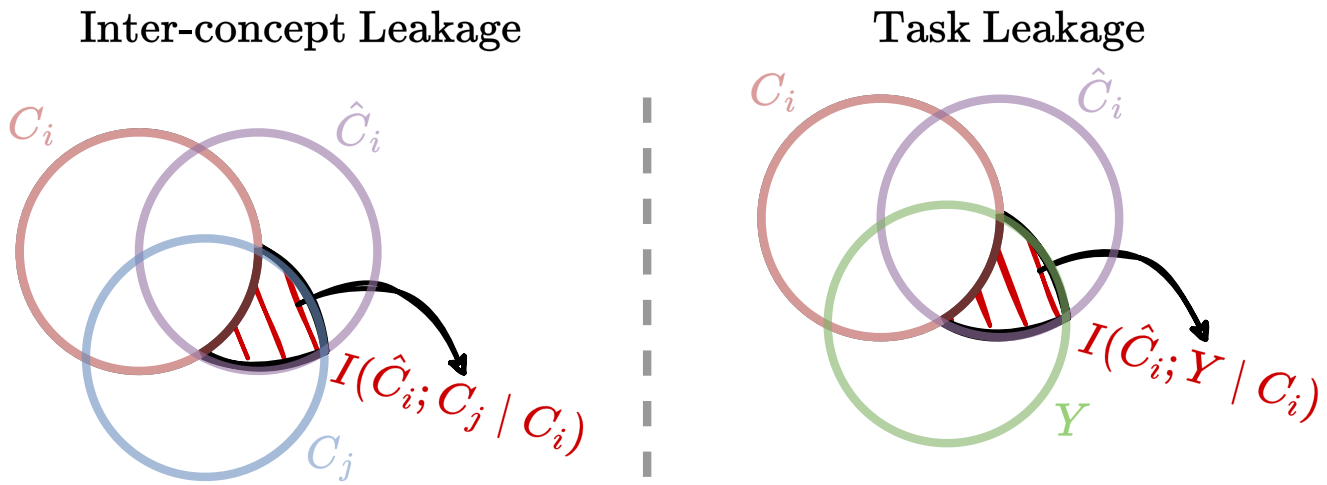


2 **Provide concept-based "hints"**



WHAT IS INFORMATION LEAKAGE?

Leakage occurs when the representation $\hat{C}_i$ learnt for a concept C_i encodes unnecessary information about other concepts C_j or the downstream task Y



THE CASE AGAINST LEAKAGE

Traditionally, the narrative is that leakage must be eradicated as it:

1. [Intervenability] Affects the intervenability of concept-based models
2. [Interpretability] Leads to less interpretable concept-based models, as evaluated by some proxy metrics for interpretability.
3. [Safety] Can lead to models learning shortcuts or privacy/fairness violations

THE CASE AGAINST LEAKAGE

Traditionally, the narrative is that leakage must be eradicated as it:

1. [Intervenability] Affects the intervenability of concept-based models
2. [Interpretability] Leads to less interpretable concept-based models, as evaluated by some proxy metrics for interpretability.
3. [Safety] Can lead to models learning shortcuts or privacy/fairness violations

I have recently developed some doubts on the first two claims.

OUR POSITION

In contrast to the common narrative about leakage, we posit that:

Not all forms of leakage are malign, and that a controlled form leakage (which we call benign leakage) can be necessary for constructing accurate and intervenable CMs

WHY WOULD WE EVER WANT LEAKAGE?

Concept sets in real-world datasets tend to be **incomplete**!

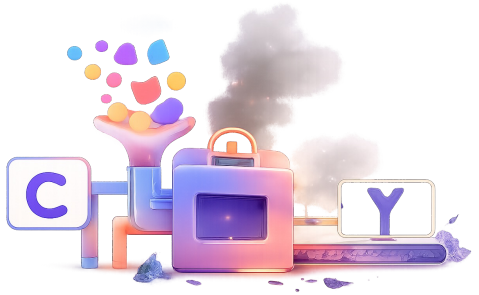
Expectation

Concept Completeness

$I(X;Y) \leq I(C;Y)$

VS

Concept Incompleteness



$I(X;Y) \gg I(C;Y)$

Reality

The only way CMs can accurately predict their downstream task when concepts are incomplete is by allowing leakage!

BENIGN LEAKAGE

Therefore, we argue that leakage should not be avoided at all costs, so long as the leakage has the following properties:

1. Sufficiency: the leakage R should all encode the information in the task Y that is not present in the set of training concepts C.
2. Localization: the representation $\hat{C}_i$ can be decomposed as $\hat{C}_i \equiv (\bar{C}_i, R_i)$, where all information about a concept C_i that is informative for predicting Y is found in $\bar{C}_i$.

We call this form of leakage **benign leakage**.

ALIGNMENT VIA INTERVENTIONS

Specifically, we show that benign leakage can be encouraged by optimizing for the model's performance after **all** concepts are intervened on:

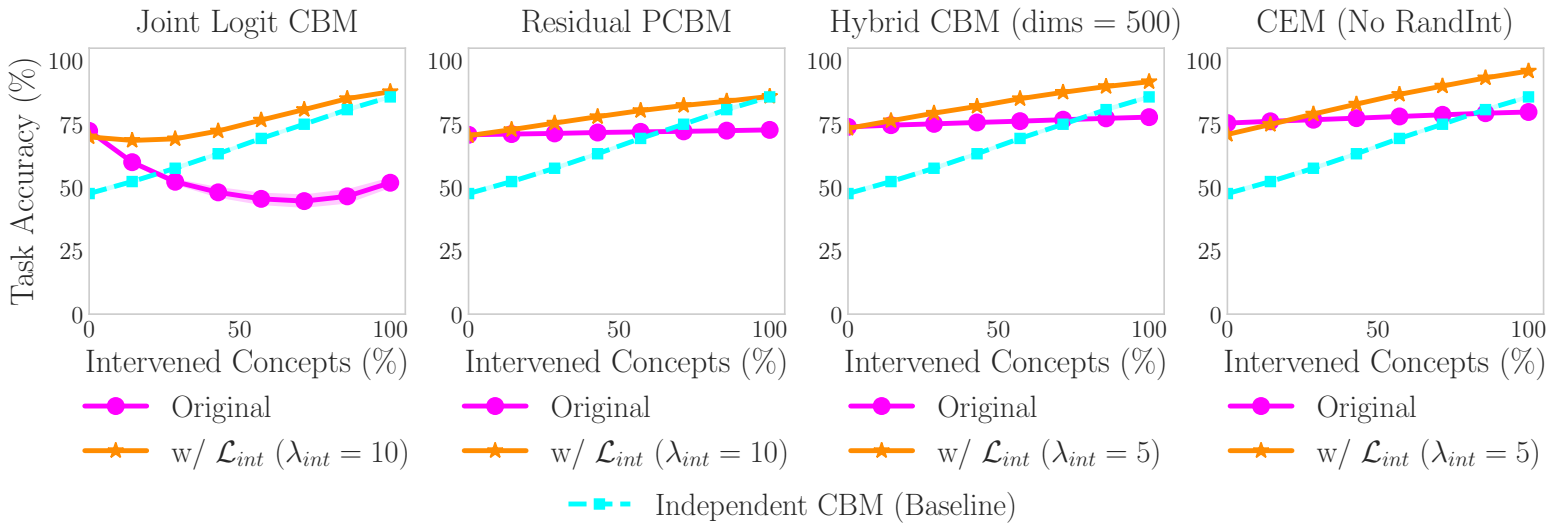
$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CBM}} + \lambda_{\text{int}} \underbrace{\mathbb{E}_{(\boldsymbol{x},\boldsymbol{c},y)\sim\mathcal{D}_{\text{train}}}\left[\text{CE}\left(f\left(g(\boldsymbol{x}; C := \boldsymbol{c})\right), y\right)\right]}_{\mathcal{L}_{\text{int}}}$$

And argue that this is something that already existing intervenable, but leaky CMs *implicitly* optimize for during training.

ALIGNMENT VIA INTERVENTIONS

When explicitly optimizing for this loss, one can learn leaky CMs that:

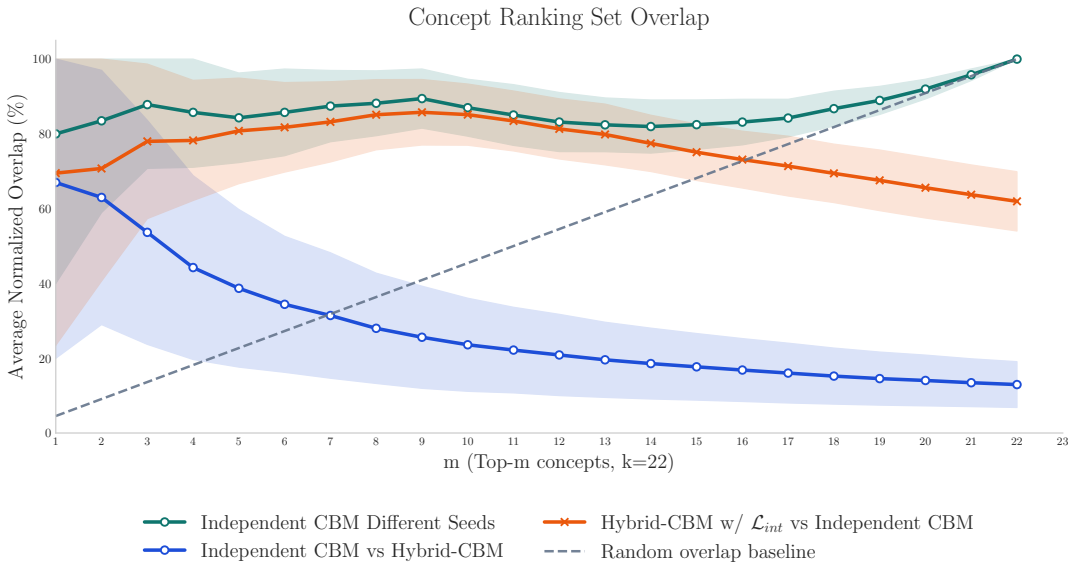
1. Are as intervenable as non-leaky CMs while remaining more accurate



ALIGNMENT VIA INTERVENTIONS

When explicitly optimizing for this loss, one can learn leaky CMs that:

1. Are as intervenable as non-leaky CMs while remaining more accurate
2. Maintain properties associated with a CM's interpretability (e.g., concept-to-task importance rankings)



ALIGNMENT VIA INTERVENTIONS

When explicitly optimizing for this loss, one can learn leaky CMs that:

1. Are as intervenable as non-leaky CMs while remaining more accurate
2. Maintain properties associated with a CM's interpretability (e.g., concept-to-task importance rankings)

These results strongly suggest that leakage does not necessarily lead to models that fail to achieve the desiderata expected from CMs

Therefore, we call for the CM research community to avoid claiming that leakage is necessarily malign, and instead explore better ways of controlling leakage

For many more results and an extended discussion, see our paper!