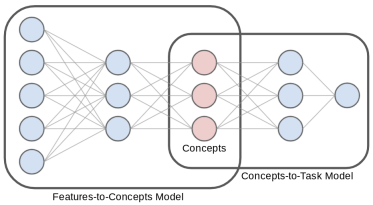
The Promises and Pitfalls of Black-Box Concept Learning Models

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, Weiwei Pan

The Promises of Concept Learning Models

Concept learning models (CLMs) help us understand how models make decisions by inserting **human-interpretable concepts**. They contain two parts: a *feature-to-concept* model and a *concept-to-task* model.



Main Contributions

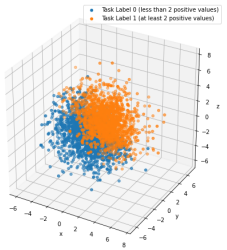
Contribution 1: We show how and when concept representations capture task-relevant info unrelated to the concepts. This informational leakage can create **misleading interpretations**.

Contribution 2: Natural mitigation techniques like adding extra dimensions, concept whitening [1], and sequential training [2, 3] do not prevent leakage.

Contribution 3: We suggest strategies to mitigate the effect of information leakage in CLMs.

Toy Task for Concept Learning

We setup a dataset of points with coordinates X, Y, and $Z \sim \mathcal{N}(0, 4)$.



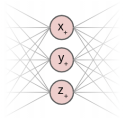
Input Features: Set of 7 non-linear, non-invertible function transformations of the coordinate values.

Concepts: x_+, y_+, and z_+ specifying if their corresponding coordinates are positive.

Labels: True (orange) if at least 2 coordinate values are positive.

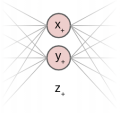
Information Leakage in CLMs

Take-away 1: Concepts are entangled in soft concept representations.



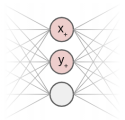
	Bottleneck Dimension		
	1	2	3
x_+	0.9999 ± 0.0001	0.65 ± 0.03	0.63 ± 0.03
y_+	0.65 ± 0.02	0.9995 ± 0.0001	0.67 ± 0.04
z_+	0.64 ± 0.02	0.66 ± 0.02	0.9995 ± 0.001

Take-away 2: Presence of unobserved concepts exacerbate information leakage.



	Bottleneck Dimension	
	1	2
x_+	0.97 ± 0.01	0.64 ± 0.01
y_+	0.66 ± 0.01	0.98 ± 0.01
z_+	0.75 ± 0.01	0.74 ± 0.01

Take-away 3: Adding extra capacity to capture unobserved concepts does not prevent leakage.



	Bottleneck Dimension		
	1	2	3
x_+	0.99998 ± 0.00001	0.59 ± 0.02	0.5 ± 0.2
y_+	0.60 ± 0.01	0.99973 ± 0.00002	0.5 ± 0.3
z_+	0.68 ± 0.02	0.693 ± 0.005	0.4 ± 0.4

Tables present AUC scores when predicting concepts from bottleneck node activations.

Concept Whitening is Not Enough

Concept Whitening [1] decorrelates soft concept representations.

Task: Predict if a subset of MNIST images (digits 1-6) display numbers smaller than 4 using an incomplete set of task relevant concepts.

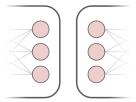
Result: Since the CW dimensions are full activation maps, we can predict all the concepts from any of the CW dimensions (aligned and unaligned) with $95 - 100\%$ accuracy.

Take-away 4: Concept Whitening does not prevent information leakage.

Task-Blind Training is Not Enough

Take-away 5: Task-blind training of concept representations does not prevent leakage.

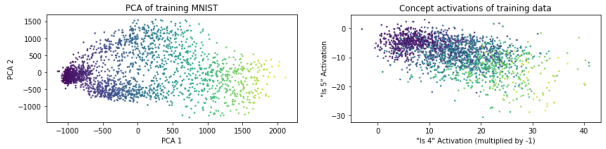
Task: Predict parity of MNIST images using ***task-irrelevant concepts***. Expect **50%** accuracy.



Model: First train pixels-to-concepts model for concept representations. Then train concepts-to-parity model. Representations are not aware of task.

Result: Model achieves **69%** on test set – much greater than 50%.

Cause: Concept representations, though task-unaware, encode first PCA dimension of MNIST data.



Furthermore

- Information leakage is sensitive to modeling choices.
- Even random concepts can improve task performance, so distinguishing between task-related concepts and random concepts can be difficult.
- Mitigation: Minimize mutual information
- Mitigation: Use human experts to refine concepts

Acknowledgements

WP was funded by the Harvard Institute of Applied Computation Science. IL was funded by NSF GRFP (grantno. DGE1745303). FDV was supported by NSF CAREER 1750358.

References

1. Chen, Z., Bei, Y., and Rudin, C. Concept whitening for inter-pretable image recognition.Nature Machine Intelligence,2(12):772–782, 2020.
2. Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson,E., Kim, B., and Liang, P. Concept bottleneck models.InInternational Conference on Machine Learning, pp.5338–5348. PMLR, 2020.
3. Margeloiu, A., Ashman, M., Bhatt, U., Chen, Y., Jamnik,M., and Weller, A. Do concept bottleneck models learn as intended? arXiv preprint arXiv:2105.04289, 2021.