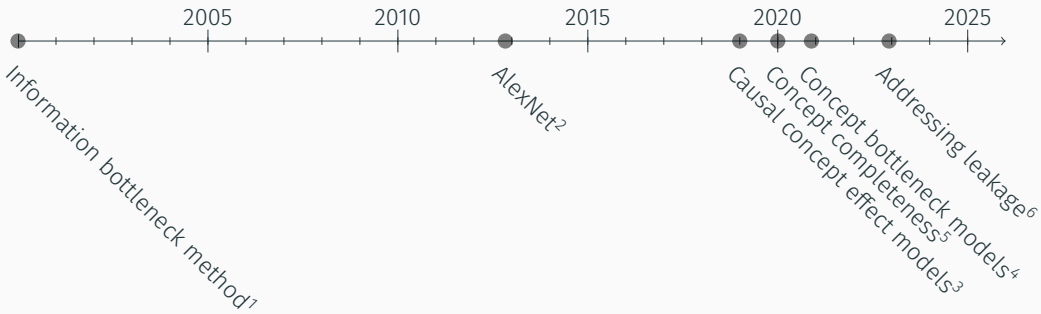
Addressing Leakage in Concept Bottleneck Models

Marton Havasi, Sonali Parbhoo, Finale Doshi-Velez

Edmund Goodman

February 21, 2025

Background and related work



[1] Tishby, Pereira, and Bialek, *The Information Bottleneck Method*.
[2] Krizhevsky, Sutskever, and Hinton, "ImageNet Classification with Deep Convolutional Neural Networks".
[3] Goyal et al., *Explaining Classifiers with Causal Concept Effect (CaCE)*.
[4] Koh et al., "Concept Bottleneck Models".
[5] Yeh et al., "On Completeness-aware Concept-Based Explanations in Deep Neural Networks".
[6] Havasi, Parbhoo, and Doshi-Velez, "Addressing Leakage in Concept Bottleneck Models".

Concept bottleneck models (CBMs)[7] recap

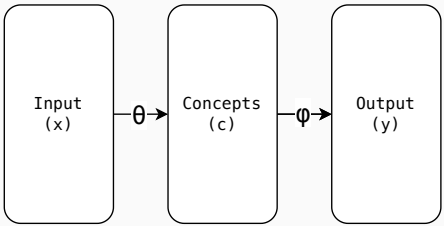


Figure 1: A diagram of a CBM. Given a dataset $\mathcal{D} = \{x, c, y\}$ with inputs $x \in \mathbb{R}^d$, K binary concepts $c \in \{0, 1\}^K$, and output labels $y \in \mathbb{N}$, a concept predictor θ and label φ can be trained.

Hard CBMs

- Label predictor takes binary concept values
- Concept and label predictors trained separately on ground truth

Soft CBMs

- Label predictor takes real concept probabilities
- Can be trained independently, sequentially, or jointly

[7] Koh et al., "Concept Bottleneck Models".

Where does this fit in?

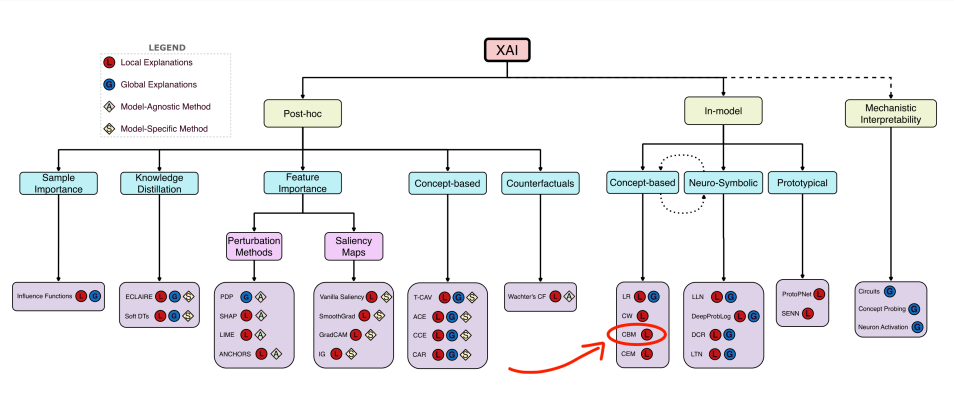


Figure 2: The CBM approach within the taxonomy of explainable AI, from the lecture slides[8].

[8] Jamnik, Shams, and Zarlenga, "Explainable Artificial Intelligence (L193), Lecture 1, Lent 2025".

Leakage

- "The label predictor learns to utilize the **additional, unintended information** in the soft concept probabilities output by the concept predictor"[9]
- Learnt during sequential or joint training of soft CBMs
  - Need to be able to "see" input values x_i during training of φ
- Damages both interpretability and interventions
  1. Label explanation no longer wholly based on concept values
  2. Intervening on a concept may not result in correct labelling

[9] Havasi, Parbhoo, and Doshi-Velez, "Addressing Leakage in Concept Bottleneck Models", §2, Leakage.

Trade-off between accuracy and leakage

- Hard CBMs have no leakage, but are less accurate
- Soft CBMs have better accuracy, but leakage damages interpretability and interventions

- *Can we get the best of both worlds?*

Key idea

Two key factors cause the disparity between the accuracy of hard and soft CBMs

1. The Markovian assumption
2. The expressivity of concept predictors

Inherent to CBMs, but side-stepped by soft approaches leaking information – at a cost...

The Markovian assumption

Hard CBMs assume that the concepts capture enough information about the input to label the output

- If not enough concepts, this may not be true!
  - Cannot distinguish between cats and dogs if the only concepts are having fur and having a tail...
- More formally, there is no mutual information shared between y and x given c, written as $I(y; x|c) = 0$

Addressing the Markovian assumption

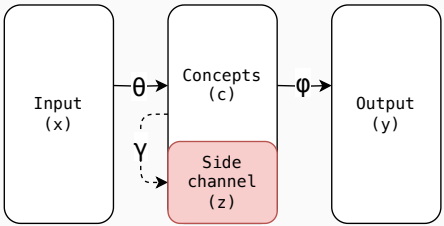


Figure 3: A diagram of a CBM with a side-channel. Introduce L unknown concepts $z \in \{0,1\}^L$, which can capture required information about x not in the existing concepts c. γ is a learnt function to infer z to c

- Consolidate leakage into separate unknown concepts
- Training similar to hard CBM, but modified to estimate gradients for unknown concepts
- Completeness score $\frac{I(y;c)}{I(y;c,x)}$ estimates what fraction of the required information is present in c
- If the label must be fully explained, use γ to predict z from c

  $\Rightarrow$ **Intuition: leakage only during training!**

The expressivity of concept predictors

All CBMs assume independence of concepts, so cannot capture correlations between them

- Trivial example of failure case is mutually exclusive concepts
  - Labelling an animal reported as having both toes and hooves is meaningless...

Addressing the expressivity of concept predictors

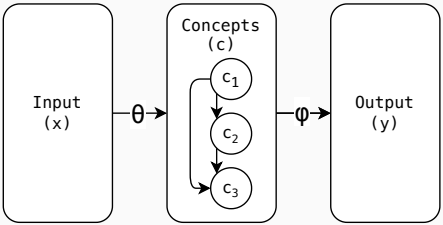


Figure 4: A diagram of a CBM with an auto-regressive architecture.

- To predict the i^{th} concept, use not only the input x, but also the already predicted concepts $c_{1:i1}$
  - In Figure 4, c_2 depends on c_1 and c_3 depends on both c_1 and c_2
- Allows correlations between concepts to be captured
- Again, training similar to hard CBM, but more involved modification required
  - Relegated to the paper's appendix...

Evaluation i

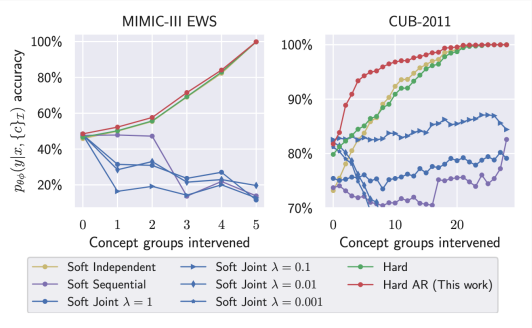


Figure 5: Label accuracy after intervening on subset of concept[a].

[a] Havasi, Parbhoo, and Doshi-Velez, "Addressing Leakage in Concept Bottleneck Models", Figure 2.

- Tested for both prediction (*MIMIC-III EWS*) and classification (*Caltech-UCSD Birds*) tasks
- Both side-channel and auto-regressive approach close gap in accuracy between hard and soft CBMs, but do not damage interventions
  - Figure 5 shows accuracy scaling well with interventions for both datasets
  - Figure 6 shows side-channel closing accuracy gap
- Significant increase to computation cost

Evaluation ii

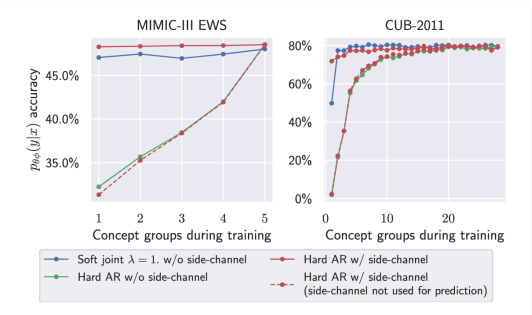


Figure 6: The predictive performance of soft joint CBMs and hard CBMs[a].

[a] Havasi, Parbhoo, and Doshi-Velez, "Addressing Leakage in Concept Bottleneck Models", Figure 3a.

- Tested for both prediction (*MIMIC-III EWS*) and classification (*Caltech-UCSD Birds*) tasks
- Both side-channel and auto-regressive approach close gap in accuracy between hard and soft CBMs, but do not damage interventions
  - Figure 5 shows accuracy scaling well with interventions for both datasets
  - Figure 6 shows side-channel closing accuracy gap
- Significant increase to computation cost

Summary

Get the best of both worlds between hard and soft CBM accuracy and interpretability

⇒ Leakage damages interpretability and interventions in soft CBMs

⇒ Hard CBMs have two causes for accuracy shortcomings

⇒ Using a side-channel and auto-regressive architecture can fix these issues

Critique

Strengths:

+ Approach follows clear train of logic, identifying and addressing issues
+ Strong motivation, supported by empirical measurement
+ Neat trick using γ function to leak only during training
+ Open source code[10] strengthens argument and replicability

Weaknesses:

– Side-channel consolidates uninterpretable leakage rather than removing it, with a solution left for future work[11]
– Significant and variable computational cost (from 1.6× to 11×) may impact scalability

[10] Actionable Knowledge (DtAK) Lab, *Dtak/Addressing-Leakage*.
[11] Angluin, "Proof Techniques".

Impact and future work

- Commonly cited (74 citations) as related work on concept embedding models by eminent researchers...[12]
- Few papers directly building on the approach
- Critique/future work of side-channels being difficult to interpret is supported by "Benchmarking and Enhancing Disentanglement in Concept-Residual Models"[13]

[12] Zarlenga et al., "Learning to Receive Help".
[13] Zabounidis et al., *Benchmarking and Enhancing Disentanglement in Concept-Residual Models.*

Addressing Leakage in Concept Bottleneck Models

Marton Havasi, Sonali Parbhoo, Finale Doshi-Velez

Edmund Goodman

February 21, 2025

References i

[1] Data to Actionable Knowledge (DtAK) Lab. ***Dtak/Addressing-Leakage.*** Data to Actionable Knowledge (DtAK) Lab, Nov. 7, 2024. URL: https://github.com/dtak/addressing-leakage (visited on 02/15/2025).

[2] Dan Angluin. **"Proof Techniques".** In: *SIGACT News* 15 (Winter-Spring 1983 #1 1983). URL: https://mfleck.cs.illinois.edu/proof.html (visited on 02/15/2025).

[3] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. ***Explaining Classifiers with Causal Concept Effect (CaCE).*** Feb. 28, 2020. DOI: 10.48550/arXiv.1907.07165. arXiv: 1907.07165 [cs]. URL: http://arxiv.org/abs/1907.07165 (visited on 02/15/2025). Pre-published.

References ii

[4] Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. **“Addressing Leakage in Concept Bottleneck Models”.** In: *Advances in Neural Information Processing Systems* 35 (Dec. 6, 2022), pp. 23386–23397. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/944ecf65a46feb578a43abfd5cddd960-Abstract-Conference.html (visited on 02/15/2025).

[5] Mateja Jamnik, Zohreh Shams, and Mateo Espinosa Zarlenga. **“Explainable Artificial Intelligence (L193), Lecture 1, Lent 2025”.** Jan. 24, 2025. URL: https://www.cl.cam.ac.uk/teaching/2425/L193/files/Cambridge-XAI-MPhil-L193-2025-Lecture1.pdf (visited on 02/17/2025).

References iii

[6] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. **"Concept Bottleneck Models".** In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, Nov. 21, 2020, pp. 5338–5348. URL: https://proceedings.mlr.press/v119/koh20a.html (visited on 02/15/2025).

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. **"ImageNet Classification with Deep Convolutional Neural Networks".** In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. URL: https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (visited on 11/12/2024).

References iv

[8] Naftali Tishby, Fernando C. Pereira, and William Bialek. ***The Information Bottleneck Method.*** Apr. 24, 2000. DOI: 10.48550/arXiv.physics/0004057. arXiv: physics/0004057. URL: http://arxiv.org/abs/physics/0004057 (visited on 02/15/2025). Pre-published.

[9] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. **"On Completeness-aware Concept-Based Explanations in Deep Neural Networks".** In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 20554–20565. URL: https://proceedings.neurips.cc/paper/2020/hash/ecb287ff763c169694f682af52c1f309-Abstract.html (visited on 02/18/2025).

References v

[10] Renos Zabounidis, Ini Oguntola, Konghao Zhao, Joseph Campbell, Simon Stepputtis, and Katia Sycara. ***Benchmarking and Enhancing Disentanglement in Concept-Residual Models.*** Nov. 30, 2023. DOI: 10.48550/arXiv.2312.00192. arXiv: 2312.00192 [cs]. URL: http://arxiv.org/abs/2312.00192 (visited on 02/17/2025). Pre-published.

[11] Mateo Espinosa Zarlenga, Katie Collins, Krishnamurthy Dvijotham, Zohreh Shams, and Mateja Jamnik. **"Learning to Receive Help: Intervention-Aware Concept Embedding Models".** In: *Advances in Neural Information Processing Systems* 36 (Dec. 15, 2023), pp. 37849–37875. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/770cabd044c4eacb6dc5924d9a686dce-Abstract-Conference.html (visited on 02/17/2025).